Inference from estimates of exposure effects using stratification on the propensity score

Elizabeth Williamson

A thesis submitted for the degree of Doctor of Philosophy of the University of London

London School of Hygiene & Tropical Medicine May 2007

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Statement of Own Work

All students are required to complete the following declaration when submitting their thesis. A shortened version of the School's definition of Plagiarism and Cheating is as follows (the full definition is given in the Research Degrees Handbook):

The following definition of plagiarism will be used:

Plagiarism is the act of presenting the ideas or discoveries of another as one's own. To copy sentences, phrases or even striking expressions without acknowledgement in a manner which may deceive the reader as to the source is plagiarism. Where such copying or close paraphrase has occurred the mere mention of the source in a biography will not be deemed sufficient acknowledgement; in each instance, it must be referred specifically to its source. Verbatim quotations must be directly acknowledged, either in inverted commas or by indenting. (University of Kent)

Plagiarism may include collusion with another student, or the unacknowledged use of a fellow student's work with or without their knowledge and consent. Similarly, the direct copying by students of their own original writings qualifies as plagiarism if the fact that the work has been or is to be presented elsewhere is not clearly stated.

Cheating is similar to plagiarism, but more serious. Cheating means submitting another student's work, knowledge or ideas, while pretending that they are your own, for formal assessment or evaluation.

Supervisors should be consulted if there are any doubts about what is permissible.

Declaration by Candidate

I have read and understood the School's definition of plagiarism and cheating given in the Research Degrees Handbook. I declare that this thesis is my own work, and that I have acknowledged all results and quotations from the published or unpublished work of other people.

Signed:..... [redacted] ..... Date:20/05/07.....

Full name:.... [redacted] ..... (please print clearly)

Acknowledgements

I am very grateful to my supervisor, Dr. James Carpenter, for his support, encouragement and excellent advice. I would also like to thank the staff and students of LSHTM who gave me ideas and inspiration, in particular Professor Stephen Evans, Professor Stuart Pocock and Dr. Joe Kim. I am grateful to Dr. Stijn Vansteelandt at Ghent university, and the staff at Freiburg university who gave me very helpful feedback on my thesis. I would also like to mention my fellow PhD students, without whose positivity, enthusiasm and company I could not have managed.

Chapter 7 uses data from the ESCAPE trial from King's College, London, kindly provided by Dr. Mike Hurley.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Many observational epidemiological studies are concerned with estimating the causal effect of a treatment or exposure, hereafter referred to as the treatment. In addition to substantial philosophical controversy surrounding the subject of causality [41] there are also considerable methodological difficulties in estimating causal treatment effects from observational data. In a randomised study, the randomisation leads us to expect, on average, that the treatment groups are comparable in all characteristics affecting the outcome other than treatment status. This comparability allows unbiased estimation of the causal treatment effect [37]. Conversely, in observational studies there are usually systematic differences in the characteristics of subjects between treatment groups. If these characteristics are related to the outcome then estimates of the causal treatment effect will be biased — a problem which epidemiologists refer to as confounding [110].

Despite the difficulties associated with estimating causal treatment effects in the presence of confounding, observational studies continue to be used to investigate causal epidemiological questions. This is because randomised trials are often unfeasible, due to, for example, ethical, financial or practical reasons [10]. In these situations, we must rely on observational studies to estimate the causal effect of a treatment. It is therefore important to be able to tackle causal questions in the presence of confounding.

Methods of dealing with confounding in observational studies can be split, rather crudely, into two categories — design-based and model-based. The design-based methods attempt to define classes of subjects within which subjects in different treatment groups are comparable in all characteristics affecting the outcome other than treatment status. Each class of subjects then mimics a randomised study, allowing unbiased estimation of the causal treatment effect within that class. These methods

crucially rely on our ability to define such classes. The model-based methods, on the other hand, posit some causal relationship between treatment status, subject characteristics and the outcome, and use this structure to estimate the causal treatment effect. These methods produce unbiased estimates of causal treatment effects only when the model is sufficiently true to life.

We now consider some types of design-based methods for dealing with confounding. A method popular with epidemiologists is stratification on the observed confounding variables, which includes standardization [25] and the Mantel-Haenszel methods [62]. Matching on the observed confounding variables is also a popular method [68]. The matched pairs or strata, however, will only mimic a randomised study, in terms of comparability of treatment groups within each matched pair or strata, when it is possible to stratify exactly by each confounding variable. When the number of confounding variables is large this becomes unfeasible. As a solution to this problem, propensity scores methods were proposed [84]. Provided that all confounding variables are observed, stratifying or matching on the propensity score can produce unbiased estimates of causal treatment effects. This, since the propensity score is a scalar variable, is much easier than stratifying or matching simultaneously on many variables. All these methods, however, share one important disadvantage: they cannot adjust for confounding by unobserved variables. In order to overcome this problem, instrumental variables methods can be applied [64]. If a suitable 'instrument' can be found — a variable that is correlated with treatment status but independent of all confounding variables — then both observed and unobserved confounding variables are dealt with. However, it has been noted that it is often difficult to find an instrument in epidemiological studies when the confounding is severe [64].

We briefly mention some model-based methods for dealing with confounding. In epidemiology, the most popular methods of this type are maximum likelihood regression models [52]. It has been shown, however, that if the mathematical assumptions implicit in these models are violated, regression can produce biased estimates of causal treatment effects [88]. Structural equation models attempt to move beyond modelling merely the association between treatment status, subject characteristics and the outcome, by proposing a model for the within-subject causal relationship between treatment and outcome, specifying the way in which confounding variables interrelate [31]. Again, the results are crucially dependent on the structural assumptions made. Directed acyclic graph (DAG) methods can alleviate this problem by

making the causal assumptions explicit, and can be used to check whether the observed variables are sufficient and appropriate to control for [73]. However, DAGs do not provide a means of testing the causal assumptions made.

Of all these methods for dealing with confounding in observational studies, this thesis focuses on propensity score methods. The reason for this choice is that since a landmark paper introducing propensity scores in 1983 [84], their use in epidemiological applications has increased greatly each year [102]. As we will see, however, it is not clear how well propensity score methods perform in comparison with maximum likelihood regression models, nor is there much guidance about which of the various propensity score methods should be used. In this thesis, we focus on the method of stratification on the propensity score. In particular, we consider the issue of making inferences from the resulting estimator, which we call the stratified treatment effect estimator. The first aim of this thesis, therefore, is to ascertain the large-sample properties of the stratified treatment effect estimator from a frequentist perspective — consistency, the asymptotic sampling distribution, and the asymptotic variance. The second aim is to investigate methods of constructing confidence intervals for the stratified treatment effect estimator. These aims have a two-fold purpose: to facilitate the practical application of stratification on the propensity score, and to add to the growing methodological literature about propensity score methods in order that fair comparisons can be made between different propensity score methods and the standard regression models.

The use of propensity score methods is motivated through a randomisation argument in Chapter 2, where we show that unbiased estimates of causal treatment effects can be obtained when confounding is present by adjusting for the propensity score. We then describe four main propensity score methods in detail and apply them to an artificial dataset. We review published comparisons of the various propensity score methods, and attempt to draw links between them in order to more clearly understand the relative merits of each.

In Chapter 3 we derive the large-sample properties of the stratified treatment effect estimator. In particular, we ascertain conditions under which it is consistent and asymptotically normally distributed. We calculate its asymptotic variance, assuming that the propensity score is: (i) a known function of the observed covariates, and (ii) estimated using a correctly specified logistic regression model. These variances

are denoted by $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$, respectively, and are expressed in terms of four variance components of which only the first has previously been derived. We then discuss the source of error measured by each of these components.

We begin Chapter 4 by calculating the variance of the stratified treatment effect estimator conditional on treatment status and the observed covariates. Assuming that the propensity score is a known function of the observed covariates, we then marginalise this conditional variance over the distribution of the treatment and observed covariates, using first-order approximations, obtaining the variance calculated previously, $\mathbb{V}_k[\hat{\beta}^s]$. In this way, we see that $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$ are asymptotic marginal variances of the stratified treatment effect estimator.

In Chapter 5 we calculate the four variance components contained in $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$ for a simple hypothetical dataset. We vary the example parameters one at a time in order to see if the change in the four variance components accords with our intuition gained through a discussion of the mathematical meaning of these four variance components. We then proceed to investigate the convergence rate of the two variance formulæ, by comparing the calculated values of $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$ with empirical estimates of the same variances, obtained using various sample sizes.

In Chapter 6 we consider the estimation of the variances $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$ from a sample dataset. We use kernel density estimation methods to estimate these variances. We then use these variance estimators to construct confidence intervals for a simulated dataset.

In Chapter 7, we apply the methods developed in this thesis to an observational subset of data obtained from a randomised controlled trial of an exercise program aimed at alleviating knee pain in the elderly. We use this dataset to investigate the effect of a non-randomised exposure that was observed during the trial.

We end, in Chapter 8, by summarizing and discussing our results. Practical guidance for epidemiologists arising from the work in this thesis and suggestions about when the methods developed here should be used are given. Potential extensions of this work and other promising avenues of research in this area are also discussed.

Chapter 2

Review of propensity score methods

In order to place the current research in context, we begin by reviewing the propensity score literature. We first explain the theoretical justification for the use of propensity score methods to estimate causal effects in the presence of confounding. We then describe four propensity score methods: stratification on the propensity score, matching on the propensity score, covariate adjustment including the propensity score, and weighting by the inverse of the propensity score. The advantages and disadvantages of each method are briefly discussed and the extent of their use is reviewed, with particular emphasis on epidemiological applications.

2.1 The use of propensity score methods to estimate causal effects

We begin by setting the scene. As usual in a frequentist setting, we assume repeated-sampling from a near-infinite [1] population indexed by fixed but unknown parameters. We first consider a simple scenario, where the outcome, Y, is continuous and depends on a binary treatment [2], Z, and a set of covariates, $\mathbf{X} = (X_1, ..., X_m)$. We wish to estimate the causal effect of the treatment, Z, on the outcome, Y, from a sample of data, $\{Y_i, Z_i, \mathbf{X}_i\}$ for $i = 1, ..., n$, drawn independently from the population.

In order to clearly define the 'causal effect of the treatment', we introduce the potential outcomes framework, whose formalization is often attributed to Rubin [41], where, for a particular subject, Y_1 denotes the outcome we would have seen had that subject been treated ($Z = 1$) and Y_0 denotes the outcome we would have seen had that subject not been treated ($Z = 0$). The observed outcome, Y, can be written as $Y = Y_1 Z + Y_0 (1 - Z)$.

[1] We will usually be dealing with finite populations but we assume that these are so large that the correction is negligible.

[2] In observational epidemiological studies this will often be an exposure rather than a treatment. However, we refer to a 'treatment' throughout for consistency and brevity.

We now define the causal treatment effect using the potential outcomes notation. For a particular subject, any causal quantity can be described as a contrast between the two potential outcomes, Y_1 and Y_0. In particular, we define the causal treatment effect for an individual as $Y_1 - Y_0$, the difference between the two outcomes they could potentially have experienced. We are interested in the average causal treatment effect across the whole population. We call this the population average causal treatment effect, denoted by β_o where,

$$\beta_o = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]. \tag{2.1}$$

Although we focus on the estimation of β_o, we sometimes discuss another causal treatment effect — the population average causal treatment effect *on the treated*. We denote this by β_o^t where,

$$\beta_o^t = \mathbb{E}[Y_1 \mid Z = 1] - \mathbb{E}[Y_0 \mid Z = 1]. \tag{2.2}$$

The two estimands β_o^t and β_o will be different when the causal treatment effect for individual subjects, $Y_1 - Y_0$, depends on covariates related to treatment status. Which of these estimands we wish to estimate will depend on the question we wish to answer. For example, if we wanted to evaluate the efficacy of a flu vaccine, we would probably be interested in the effect it had on the weak and elderly — those who usually receive it. We would be less interested in estimates of the effect of vaccinating the whole population of Britain. In this case, β_o^t would be a more appropriate estimand than β_o. Conversely, suppose we were interested in estimating the effect of compulsory school meals on the obesity levels of British schoolchildren. In this case, we would want to know the effect of these healthier meals on the whole school population of Britain, rather than on the sub-population of children who are already likely to eat a healthy diet. The appropriate estimand here would be β_o. Although both β_o and β_o^t are discussed later in this chapter, we focus primarily on the estimation of β_o. Before considering particular methods of using the propensity score to estimate β_o, two standard assumptions are made.

Assumption 2.1 *The potential outcomes, covariates and treatment,* $(Y_0, Y_1, \mathbf{X}, Z)$*, are independently and identically distributed for each subject. Specifically, the distribution of the potential outcomes for one subject is independent of the treatment status of another subject, given the observed covariates.*

The second half of Assumption 2.1 has been called the Stable Unit Treatment Value Assumption (SUTVA) [90]. A nice example of a violation of SUTVA is given by Little and Rubin [57], which is as follows. Suppose you and I are in the same room, both with headaches. Your taking aspirin will affect the state of my headache whether or not I take aspirin since if you don't take aspirin, your whinging will counteract any alleviating effect of my aspirin!

Assumption 2.2 *Treatment assignment and the potential outcomes are conditionally independent, given the observed covariates,* $\mathbf{X}$. *Mathematically,* $\{ Z \perp (Y_0, Y_1) \} \mid \mathbf{X}$, *where* $\perp$ *is used to denote conditional independence [21].*

This assumption is frequently termed strongly ignorable treatment assignment (given the observed covariates) [84]. It has also been called selection on observables [36], and merely states that there are no unobserved confounders. In a randomised study, we can expect this to be true even when no covariates are observed. In observational studies, since there is no statistical test of this assumption we must use our knowledge of the problem and the data collected in order to judge how plausible it is that all confounders have been observed.

We now consider how, under Assumption 2.1, propensity score methods use Assumption 2.2 to estimate causal treatment effects from a sample of data when confounding is present.

2.1.1 Adjusting for confounding using the observed covariates

We seek to estimate the population average causal treatment effect, β_o, from a sample of data. A naive way to estimate this would be to take the difference in mean outcomes of treated and untreated sampled subjects. This estimates

$$\mathbb{E}[Y_1 \mid Z = 1] - \mathbb{E}[Y_0 \mid Z = 0]. \tag{2.3}$$

In the absence of confounding, on average the treated and untreated groups are comparable in terms of all characteristics that affect the outcome, other than treatment status. Then $\mathbb{E}[Y_1 \mid Z = 1] = \mathbb{E}[Y_1]$ and $\mathbb{E}[Y_0 \mid Z = 0] = \mathbb{E}[Y_0]$. Therefore, if there is no confounding, as in a randomised study, the difference in mean outcomes of treated and untreated sampled subjects is an unbiased estimate of β_o.

In observational epidemiological data, confounding is invariably present. In this case, (2.3) is not equal to β_o, so the difference in mean outcomes of treated and untreated sampled subjects is a biased estimate of β_o. If, however, as in Assumption 2.2, treatment is strongly ignorable given the observed covariates, then of those sampled subjects whose covariate values are $\mathbf{X} = \mathbf{x}$, on average the treated and untreated groups are comparable in terms of all characteristics that affect the outcome, other than treatment status. Therefore, the difference in mean outcomes of treated and untreated sampled subjects whose observed covariate values are $\mathbf{X} = \mathbf{x}$ estimates

$$\mathbb{E}[Y_1 \,|\, Z = 1, \mathbf{X} = \mathbf{x}] - \mathbb{E}[Y_0 \,|\, Z = 0, \mathbf{X} = \mathbf{x}] \;=\; \mathbb{E}[Y_1 - Y_0 \,|\, \mathbf{X} = \mathbf{x}].$$

In this way, an unbiased estimate of treatment effect can be obtained at each observed value of the covariates. It follows that an unbiased estimate of β_o can be obtained by averaging these estimates over the distribution of the observed covariates, since

$$\beta_o = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] \;=\; \mathbb{E}_{\mathbf{X}}[\mathbb{E}[Y_1 - Y_0 \,|\, \mathbf{X} = \mathbf{x}]].$$

An analogous argument is now used to justify the use of propensity score methods to estimate causal treatment effects when confounding is present.

2.1.2 Adjusting for confounding using the propensity score

The propensity score was popularized by Rosenbaum and Rubin [84] and is defined as the conditional probability of receiving treatment given the observed covariates, which we write as $p(\mathbf{X}) = \mathbb{P}(Z = 1 \,|\, \mathbf{X})$. This score is assumed to be bounded away from zero and one, so each subject has a non-zero probability of being in either treatment group. Rosenbaum and Rubin showed that the propensity score is a 'balancing score' — in other words, that at any value of the propensity score, the population covariate distributions of treated and untreated subjects are the same, so that $\{\mathbf{X} \perp Z\} \,|\, p(\mathbf{X})$. The key idea, for causal inference, is that if treatment assignment is strongly ignorable given the observed covariates, then this balancing property of the propensity score implies that treatment assignment is strongly ignorable given the propensity score [84]. Mathematically,

$$\{Z \perp (Y_0, Y_1)\} \,|\, \mathbf{X} \qquad \Rightarrow \qquad \{Z \perp (Y_0, Y_1)\} \,|\, p(\mathbf{X}).$$

If treatment is strongly ignorable given the propensity score, then of those sampled subjects whose propensity score value is $p(\mathbf{X}) = p$, on average the treated and un-

treated groups are comparable in terms of all characteristics that affect outcome, other than treatment status. So if Assumption 2.2 is satisfied, we have a pseudo-randomised study at each value of the propensity score. Then the difference in mean outcomes of treated and untreated sampled subjects who have a propensity score of $p(\mathbf{X}) = p$ estimates

$$\mathbb{E}[Y_1 \mid Z = 1, p(\mathbf{X}) = p] - \mathbb{E}[Y_0 \mid Z = 0, p(\mathbf{X}) = p] = \mathbb{E}[Y_1 - Y_0 \mid p(\mathbf{X}) = p].$$

In this way, an unbiased estimate of the treatment effect at each value of the propensity score can be obtained by taking the difference in mean outcomes of treated and untreated sampled subjects who have that value of the propensity score. It follows that an unbiased estimate of β_o can be obtained by averaging these estimates over the distribution of the propensity score, since

$$\beta_o = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \mathbb{E}_{p(\mathbf{X})}[\mathbb{E}[Y_1 - Y_0 \mid p(\mathbf{X}) = p]].$$

This randomisation-based argument justifies the use of propensity scores to estimate causal treatment effects.

It is important to note that although we have described propensity score methods as an attempt to recreate a randomised situation, there are two important differences between randomised trials and propensity score methods. The first concerns Assumption 2.2. Randomised trials will give an unbiased estimate of treatment effect even when no confounders are observed. Propensity score methods can only give unbiased estimates of treatment effect when all confounders are observed. The second difference is the 'large-sample' aspect of propensity score methods [114]. In a randomised study, although randomisation leads us to expect the treated and untreated groups to be comparable in terms of all characteristics that affect the outcome, other than treatment status, there may be imbalance due to 'bad luck'. A large sample size decreases the chance of extreme imbalance. In the same way, under Assumption 2.2, we expect the treated and untreated groups to be comparable in terms of all characteristics that affect outcome, other than treatment status, at each value of the propensity score. A large sample size *at each value of the propensity score* decreases the chance of large imbalance. Zhao likens this to having a mini randomised study at each value of the propensity score, with the quality of the overall estimate of causal treatment effect depending on the quality of each of these mini randomised studies [114].

Estimating the propensity score

The argument above shows that unbiased estimates of causal treatment effects can be obtained by adjusting for the propensity score. Of course, in practice the propensity score is invariably unknown and must be estimated from the data. Rosenbaum and Rubin suggest estimating the propensity score from a discriminant analysis or logistic regression model [84]. The former assumes that the observed covariates follow a multivariate normal distribution whereas the latter assumes they follow one of a large number of exponential family distributions. Non-parametric estimators of the propensity score have also been proposed [40]. In epidemiological applications, the propensity score is typically estimated using a logistic regression model [111].

Since the implications of the estimation of the propensity score, in terms of the bias and variance of the treatment effect estimator, depend to some extent on which propensity score method is used, a discussion of these implications is left until the various propensity score methods have been described (Section 2.3.2).

2.1.3 Comparison of randomisation and propensity score methods

From the preceding discussion we know that in theory, if all confounders are observed, propensity score methods can produce unbiased estimates of causal treatment effects. The relevant question now, therefore, is whether or not they do so in real-life applications. We attempt to address this question by comparing treatment effect estimates from propensity score analyses with those from randomised trials since, as we have seen, randomisation of treatment leads us to expect an unbiased estimate of treatment effect.

In practice, randomised trials may not completely eliminate bias due to problems such as non-compliance, exclusion after randomisation and unblinding. Furthermore, due to the inclusion criteria used in a trial, the treatment effect being estimated by a randomised trial may not be the same as the treatment effect being estimated by an observational study. Thus, a difference in treatment effect estimates between randomised data and propensity score methods may not indicate that one is 'wrong'. Despite these issues, well conducted randomised studies are often considered the 'gold standard' method of obtaining unbiased estimates of causal treatment effects. Therefore, bearing in mind the above discussion, we now review studies that compare estimates of treatment effect from randomised and observational data, where the latter is analysed using propensity score methods.

In labour economics, estimates of the causal effect of a particular manpower training program obtained from both randomised and observational data, where the latter was analysed using a structural equations approach, were found to differ greatly [56] [26], sparking a debate on the worth of observational evidence, and leading some to conclude that randomised studies are the only reliable method of evaluation for such programs [5]. Propensity score methods appeared to solve the problem when an analysis of a subset of one of these observational datasets, using carefully applied propensity score methods, produced similar results to the randomised study [23], although doubt was cast on this finding when a re-analysis of the whole dataset using propensity score methods gave dissimilar estimates to the randomised study [98]. The authors of the first propensity score analysis argue that since subjects were excluded on the basis of lack of information with which to properly estimate the propensity score this disparity in causal estimates is to be expected [22]. Further work comparing randomised and observational estimates of causal effects suggests that propensity score methods tend to eliminate biases that are larger than average although they cannot be relied on to consistently produce unbiased estimates of causal effects [2, 67].

Returning to the epidemiological literature, we find a similar debate about the relative merits of observational and randomised studies [9, 15, 48, 54, 66, 75]. Two studies have addressed this issue by comparing results from a propensity score analysis of observational data with randomised evidence concerning the same clinical question. The first of these studies estimated the effects of statins in reducing all-cause mortality after acute myocardial infarction, using a clinical dataset, producing effect estimates that were comparable with randomised evidence [4]. The second investigated the causal relationship between tissue plasminogen activator on the all-cause mortality of ischemic stroke patients, using observational data from a German stroke registry [55]. Several propensity score methods were applied to the dataset, producing a wide range of estimates of causal treatment effects, contrasting markedly with the randomised evidence of no effect. After restricting the sample to subjects with a non-negligible chance of receiving the treatment — an estimated propensity score of more than 0.05 — the estimates of effect from all propensity score methods became comparable with the randomised evidence. In this sub-sample of the dataset, subjects were younger and healthier and therefore more similar to subjects who were included in the randomised study. This suggests that the disparity in causal effect estimates obtained from randomised and observational data may be, to some extent, due to

the lack of comparability in baseline characteristics of subjects included in the two types of study. In order to test this hypothesis, Tannen 'simulated' a particular randomised trial by selecting subjects from an observational dataset who satisfied the trial's inclusion criteria, were observed during the same time-frame, and who followed a similar treatment regimen [106]. He found that this observational sub-sample, when analysed by propensity score methods, gave estimates of causal treatment effect comparable with those obtained from the randomised trial.

Taking all the evidence into account, we conclude that a carefully conducted propensity score analysis, performed on a rich and accurate observational dataset, can produce estimates of causal treatment effects with small enough bias to be practically useful in real-life applications. With this conclusion, we proceed to look at four specific methods of using the propensity score to estimate causal treatment effects in the presence of confounding.

2.2 Methods of analysis using the propensity score

We now describe four particular propensity score methods in detail: stratification on the propensity score, matching on the propensity score, covariate adjustment including the propensity score, and weighting by the inverse of the propensity score. The treatment effect estimator obtained from each method is given for the simple scenario set up in Section 2.1, and applied to an artificial example dataset, which will be introduced shortly. We consider the bias and variance of each estimator and discuss proposed methods of reducing both.

As before, the population average causal treatment effect and the population average causal treatment effect on the treated are denoted by β_o and β_o^t, respectively. The treatment effect estimators obtained from the four propensity score methods are denoted by $\hat{\beta}^s$, $\hat{\beta}^m$, $\hat{\beta}^c$ and $\hat{\beta}^w$, where the hat denotes an estimator and the superscripts refer to 'stratification', 'matching', 'covariate adjustment' and 'weighting' respectively. The asymptotic expectation of these estimators — the 'true', or population values, of the estimators — are denoted by β_o^s, β_o^m, β_o^c, β_o^w, where the subscript of 'o' denotes a fixed population parameter and, as before, the superscripts refer to the analysis method used.

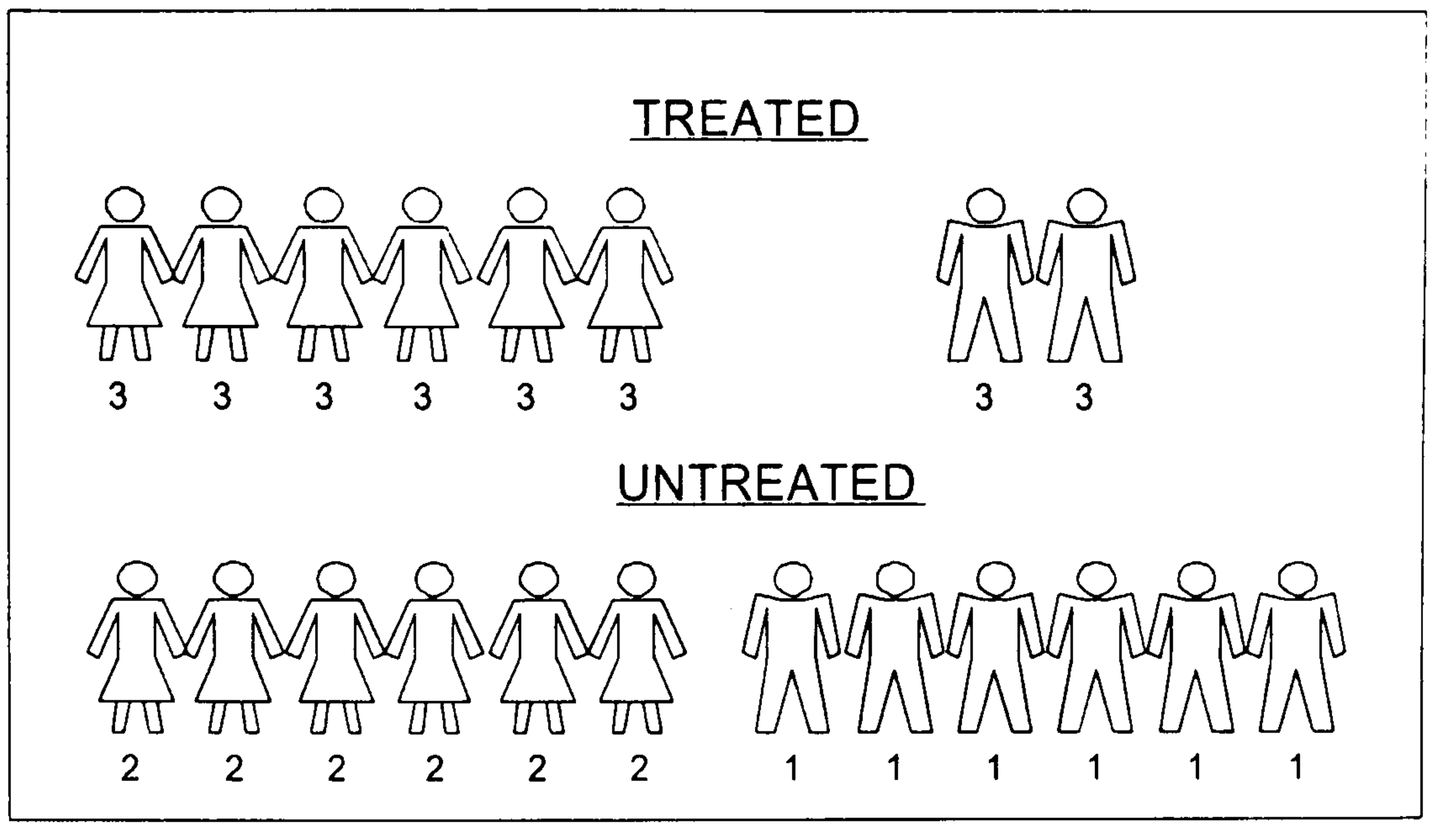
2.2.1 A hypothetical dataset

In order to see more clearly the different approaches taken by the four propensity score methods we apply each of them to an artificial example dataset, which we now describe. Suppose the minister for education wished to know whether compulsory after-school homework clubs would increase the educational achievement of British schoolchildren. In order to investigate this question the minister might collect a sample of data from a school that already runs an after-school homework club, observing a measure of educational achievement and any characteristics of the children that might impact on both their educational achievement and their attendance at the homework club. For simplicity, we assume that the only such characteristic is gender. Figure 2.1 shows an artificial dataset containing information on a sample of 20 children, independently selected from a particular school. The treatment — attendance at the homework club — is binary. The outcome is a measure of educational achievement and takes values of 1, 2 or 3. The only observed covariate, gender, is represented by figures wearing skirts and trousers, denoting girls and boys, respectively. We assume that there is no sampling variability and so this sample exactly represents the whole population. We also assume that the outcome contains no error and so gender and treatment status exactly determine the outcome value. The effect of relaxing these assumptions will be discussed in the following section.

In order to apply any of the four propensity score methods, we must make Assumptions 2.1 and 2.2 [3]. The first of these assumptions implies that the attendance of one child at the homework club will not change the effect that attendance has for another child. It is easy to think of possible violations of this assumption. For example, if a particularly badly-behaved child went to the club, he or she might disrupt everyone else and thus the educational benefit of attending the homework club would be reduced for all the other children there. The second assumption is that there are no unobserved confounders. A potential violation of this assumption is the unobserved socio-economic status of the children, since children with lower socio-economic status may be less likely to participate in after-school activities, and may also be likely to have lower educational achievement. However, at present we assume that this is not the case and that Assumptions 2.1 and 2.2 are satisfied.

[3] Note that these assumptions must also be made when a standard regression analysis is performed.

Figure 2.1: *An artificial observational dataset, where each figure represents a boy or girl. Each child's outcome is written below their figure.*



The causal treatment effects

Attendance at the homework club increases a boy's outcome by two points, and increases a girl's outcome by one point. We have assumed that there is no sampling variability and therefore the fraction of girls in the whole population is the same as the fraction of girls in the sample, which is 60%. Then the true population average causal treatment effect is

$$\beta_o \;=\; 0.4 \times 2 \;+\; 0.6 \times 1 \;=\; 1.4.$$

We also assume that the fraction of treated subjects in the whole population who are girls is the same as the fraction of treated subjects in the sample who are girls. Thus we assume that 75% of British schoolchildren who attend an after-school homework club are female. Then the population average causal treatment effect on the treated is

$$\beta_o^t \;=\; 0.25 \times 2 \;+\; 0.75 \times 1 \;=\; 1.25.$$

Since girls are more likely than boys to attend the homework club, and attendance at the homework club has less effect on girls, the population average causal treatment effect on the treated, β_o^t, is smaller than the population average causal effect for the whole population, β_o.

The estimated propensity score

Since there are both treated and untreated subjects at each combination of covariate values it is unnecessary to fit a model to estimate the propensity score. For each

combination of covariate values we estimate the propensity score by the fraction of subjects who are treated, giving

$$\mathbb{P}(\text{Treated} \mid \text{girl}) = 1/2$$
$$\mathbb{P}(\text{Treated} \mid \text{boy}) = 1/4.$$

Note that under the assumption of no sampling variability this estimated propensity score is equal to the true propensity score.

2.2.2 Stratification on the propensity score

The use of stratification (or subclassification) on the observed covariates has a long history in epidemiology [14]. Since stratification involves direct comparison of treated and untreated groups that are thought to be comparable within each stratum, it is both understandable and convincing for non-technical audiences [84]. Assumptions about the mathematical form of the outcome, and how it depends on the covariates, are not needed. When there are many covariates, however, a large number of strata must be formed in order to create strata within which all observed covariates are the same, often producing strata where all subjects have the same treatment status and so preventing the necessary within-stratum comparisons. When stratifying on the propensity score, since it is a scalar quantity, this problem is less likely to occur.

The arguments given in Section 2.1 show that exact adjustment for the propensity score can produce unbiased estimates of causal treatment effects. The propensity score, however, is often a continuous variable, in which case it is unfeasible to create strata that are exactly homogenous in the propensity score. Zhao suggests making the following assumption [114],

Assumption 2.3 *Subjects with similar propensity scores have similar covariate distributions. Mathematically, if we let $\mathbb{P}(A|B)$ refer to the conditional probability of event A given event B, for two metrics d and d', and two propensity scores, p_i and p_j, this assumption can be stated as follows. Given $\delta > 0$, there is an $\epsilon > 0$ such that*

$$d\,(p_i, p_j) < \epsilon \quad \Rightarrow \quad d'(\,\mathbb{P}(X = x \,|\, p\,(\mathbf{X}) = p_i),\ \mathbb{P}(X = x \,|\, p\,(\mathbf{X}) = p_j)) < \delta. \qquad (2.4)$$

This assumption leads us towards creating strata that are only approximately homogenous in the propensity score.

The stratified treatment effect estimator

A natural way to estimate the treatment effect would be to: (i) estimate the propensity score, (ii) split the sample into K groups using quantiles of the estimated propensity score, (iii) estimate the within-stratum treatment effects by taking the difference in mean outcome between the treated and untreated subjects in each stratum, (iv) calculate the weighted average of the within-stratum treatment effect estimates, where the weight for a particular stratum is equal to the fraction of the sample within that stratum.

Suppose we split the sample into K strata using quantiles of the estimated propensity score, where the s^{th} stratum contains a fraction r_s of the sample, and we let $\mathbf{S} = (S_1, ..., S_K)$ be a set of stratum indicators, where S_{si} is equal to one if subject i is in strata s and zero otherwise, for subjects $i = 1, ..., n$, and strata $s = 1, ..., K$. The stratified treatment effect estimator, $\hat{\beta}^s$, can be written as

$$\hat{\beta}^s = \sum_{s=1}^{K} r_s \sum_{i=1}^{n} \left\{ \frac{Y_i \, Z_i \, S_{si}}{\sum_{i=1}^{n} Z_i \, S_{si}} - \frac{Y_i \, (1 - Z_i) \, S_{si}}{\sum_{i=1}^{n} (1 - Z_i) \, S_{si}} \right\}.$$

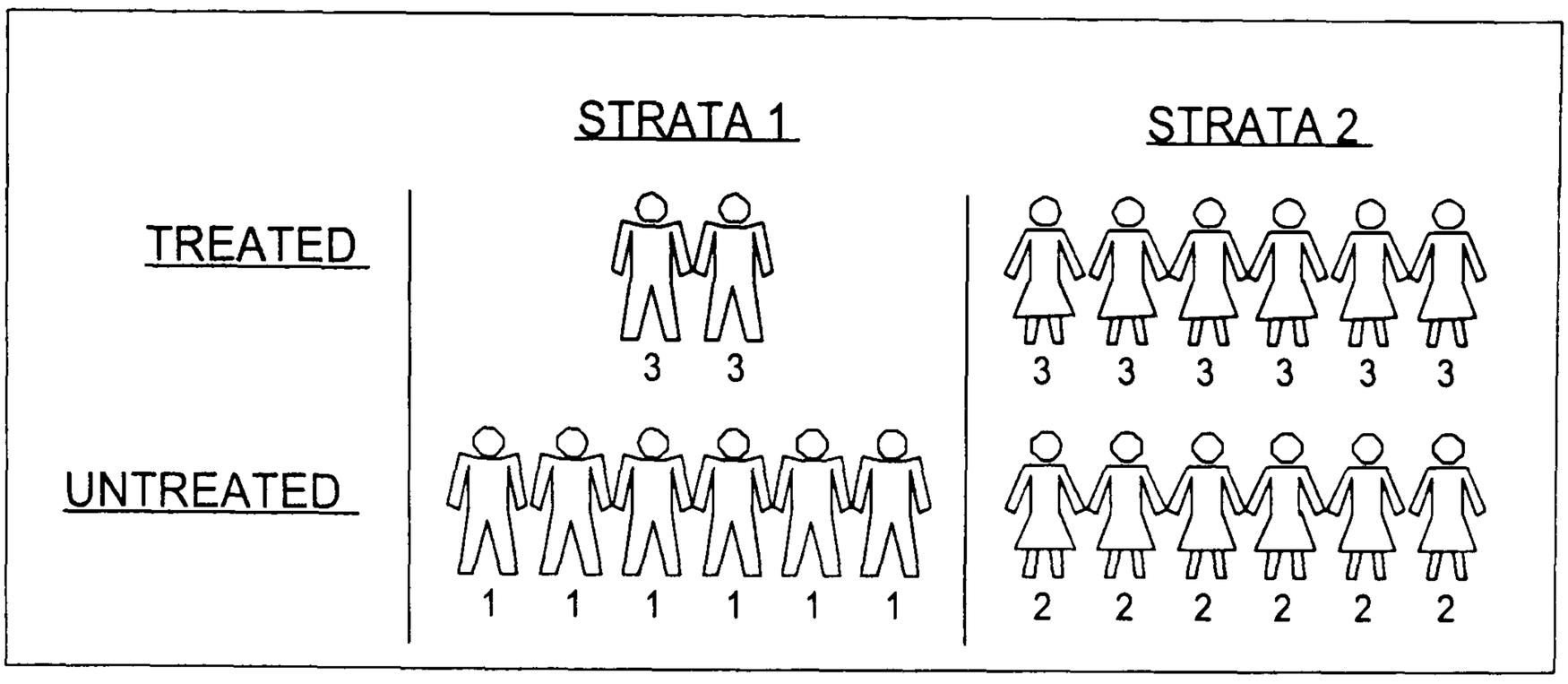
Application to the example dataset

We have already estimated the propensity score, finding that girls have an estimated propensity score of $1/2$ and boys have an estimated propensity score of $1/4$. Following the procedure outlined above, we would create two strata — one for each value of the propensity score. The two strata are shown in Figure 2.2. Given these strata, the next step is to estimate the two within-stratum treatment effects. The difference in mean outcomes of treated and untreated boys is 2, and the difference in mean outcomes of treated and untreated girls is 1. We then calculate the weighted average of these within-stratum treatment effects. There are 12 girls in a sample of 20 children so we weight the girls' treatment effect by 12/20 and the boys' treatment effect by 8/20. The stratified estimate of the effect of attendance at the homework club on educational achievement is then,

$$\hat{\beta}^s = \frac{8}{20} \times 2 + \frac{12}{20} \times 1 = 1.4.$$

Comparing this estimate with the 'true' value of β_o we see that we have correctly estimated the population average causal treatment effect.

Figure 2.2: *The stratified analysis of the dataset shown in Figure 2.1. In this example, stratifying on the estimated propensity score is equivalent to stratifying on gender.*



The bias and variance of the stratified estimator

The asymptotic expectation of $\hat{\beta}^s$, taken over our near-infinite population with all parameters fixed at their true values, is

$$\beta_o^s = \sum_{s=1}^{K} r_s \, \{\mathbb{E}[Y \mid Z = 1, S_{so} = 1] - \mathbb{E}[Y \mid Z = 0, S_{so} = 1]\},$$

where now the stratum indicators, S_{so}, rather than referring to membership of the sample strata, refer to membership of the 'true' strata, i.e. the strata that are formed by splitting the population distribution of the propensity score into K groups, with the s^{th} stratum containing a fraction r_s of the whole population, for strata $s = 1, ..., K$. Comparing β_o^s with β_o — the parameter we wish to estimate — we see that stratification on the propensity score will only produce a consistent estimator of the population average causal treatment effect if, for strata $s = 1, ..., K$,

$$\begin{aligned} \mathbb{E}[Y \mid Z = 1, S_{so} = 1] &= \mathbb{E}[Y_1 \mid S_{so} = 1], \\ \mathbb{E}[Y \mid Z = 0, S_{so} = 1] &= \mathbb{E}[Y_0 \mid S_{so} = 1]. \end{aligned} \tag{2.5}$$

We will have (2.5) if the covariate distributions within each population stratum are the same in the treated and untreated groups. In other words, the treatment must be strongly ignorable given the strata, so we must have $\{ Z \perp (Y_0, Y_1) \} \mid \mathbf{S}$, in which case β_o^s is equal to the population average causal treatment effect, β_o. When the propensity score is discrete we can create strata within which the propensity score is exactly homogenous, as in the artificial example, and so (2.5) is true. It is also true

when, for example, all the covariates associated with both treatment and outcome are constant within each stratum but other variables associated only with treatment are not constant within the stratum. In this situation, the strata would not be homogenous in the propensity score but stratification on the propensity score would still produce a consistent estimator of the population average causal treatment effect. Of course, this is not a common scenario, so we need to consider the case where treatment is not strongly ignorable given the strata. Then appealing to Assumption 2.3 we have, for strata $s = 1, ..., K$,

$$\mathbb{E}[Y_1 \mid Z = 1, S_{so}] - \mathbb{E}[Y_0 \mid Z = 0, S_{so}] \approx \mathbb{E}[Y_1 - Y_0 \mid S_{so}],$$

and so

$$\beta_o^s \approx \mathbb{E}_{\mathbf{S}}[\mathbb{E}[Y_1 - Y_0 \mid \mathbf{S}]] \equiv \beta_o.$$

Therefore, by stratifying inexactly on the propensity score, we obtain a consistent estimator of the population parameter β_o^s, which is not exactly equal to the population average causal treatment effect, β_o. However, the similarity of the propensity score within strata leads us to expect the two population estimands to be similar to one another.

Cochran shows that stratification at the quintiles of a single covariate will typically remove 90% of the bias due to that covariate when the covariate follows a number of common distributions [14], a result that has been extended to the case of stratification at the quintiles of the propensity score [85]. Cox discusses the problem of grouping data into k groups on a continuous variable, and shows that equal-sized groups are rarely optimal [16]. Equal sized strata are recommended for examples where the distribution of the stratification variable is rectangular. We will see later that the distribution of the propensity score is not usually rectangular. However, in practical applications of stratification on the propensity score, 5 equal-sized strata are typically used. We therefore adopt this choice of strata and ignore the problem of choosing the strata boundaries in the remainder of this work, although the results derived can be applied to any choice of strata boundaries.

In order to reduce the bias due to residual confounding within strata, Hullsiek and Louis compared two different methods of choosing the boundaries of the strata using simulation: choosing strata that balance the number of subjects and choosing strata that balance the inverse variance of the stratum-specific estimates of treatment effect

[44]. The latter method was found to be superior in terms of bias. An alternative approach to reducing the bias of the stratified estimator is to fit a regression model within each strata, including important predictors of outcome [59].

Although it has been suggested that bootstrapping should be used to estimate the standard error of the stratified treatment effect estimator [107], standard practice is to ignore the estimation of the propensity score and use an average within-stratum variance [59].

2.2.3 Matching on the propensity score

Matching on the observed covariates is an intuitive and transparent method of adjusting for confounding [115], which, like stratification, needs no assumptions about the mathematical form of the outcome, or its relationship with the covariates. When the observed covariate information is high-dimensional, however, finding matches for treated subjects is often impossible. Since the propensity score is a scalar variable, the problem of finding appropriate matches is greatly reduced.

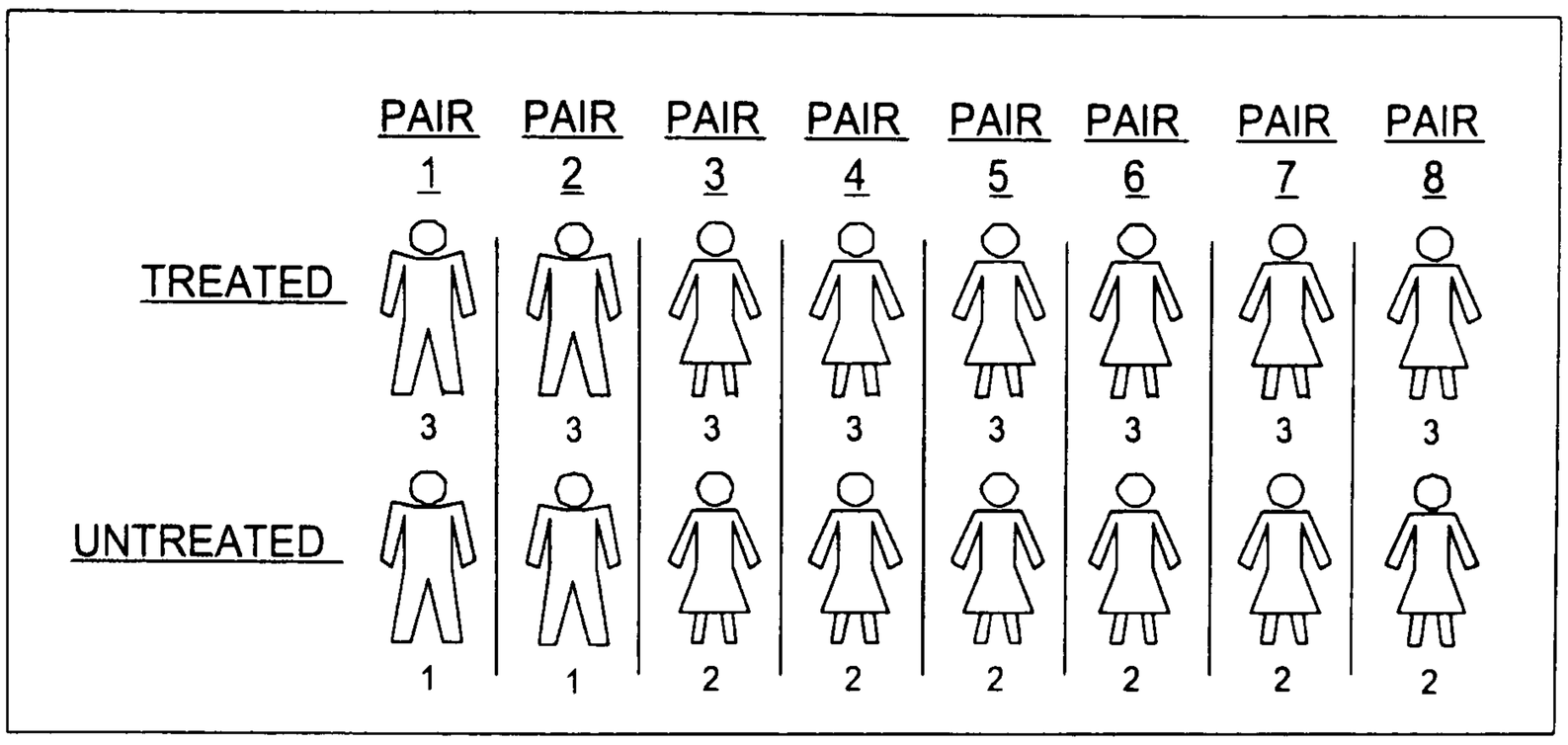
The arguments given in Section 2.1 show that exact adjustment for the propensity score can produce unbiased estimates of causal treatment effects. When continuous covariates are included in the estimation of the propensity score, however, exact matching on the propensity score may be impossible. Assumption 2.3 provides justification for inexact matching, although this will introduce some bias into the estimator.

The matched treatment effect estimator

In order to estimate a causal treatment effect using matching on the propensity score we might: (i) estimate the propensity score, (ii) for each treated subject, select a single control subject who has the same, or almost the same, value of the estimated propensity score, (iii) estimate the within-pair treatment effect by taking the difference in the two outcomes, (iv) calculate the average within-pair treatment effect estimate.

Suppose we manage to find appropriate matches for N of the treated subjects, and we let $\mathbf{M} = (M_1, ..., M_N)$ be a set of matched-pair indicators where M_{mi} is equal to one if subject i belongs to matched pair m and zero otherwise, for subjects $i = 1, ..., n$

Figure 2.3: *The matched analysis of the dataset shown in Figure 2.1. In this example, matching on the estimated propensity score is equivalent to matching on gender.*



and $m = 1, .., N$. Then the matched treatment effect estimator is $\hat{\beta}^m$, where

$$\hat{\beta}^m = \frac{1}{N} \sum_{m=1}^{N} \sum_{i=1}^{n} \{Y_i \, Z_i \, M_{mi} - Y_i \, (1 - Z_i) \, M_{mi}\}.$$

Application to the example dataset

We have already estimated the propensity score for the artificial example given in Figure 2.1, finding that the girls have an estimated propensity score of 1/2 and the boys have an estimated propensity score of 1/4. Given this estimated propensity score we match each of the two treated boys to an untreated boy, and match each of the six treated girls to an untreated girl, giving the matched pairs shown in Figure 2.3. The two matched pairs of boys have a within-pair treatment effect of 2, and the six matched pairs of girls have a within-pair treatment effect of 1. Averaging these within-pair treatment effects across the eight pairs gives the matched estimate of the effect of attendance at the homework club on educational achievement,

$$\hat{\beta}^m = \frac{1}{8} \, (2 \times 2 \; + \; 6 \times 1) = 1.25.$$

Comparing this with the 'true' population average causal treatment effects, β_o and β_o^t, we see that we have correctly estimated the population average causal treatment effect *on the treated*. In this example, since the individual-level treatment effect depends on the propensity score, β_o and β_o^t are different.

In this analysis, from a sample of 20 children, only 16 (80%) were used. Since this is an artificial situation with no error in the outcome, discarding some of the data does not change the estimate. In more realistic applications there is error in the outcome, in which case discarding data leads to an increase in the variance of the matched estimator.

The bias and variance of the matched estimator

As we might guess from the example above, the asymptotic expectation of $\hat{\beta}^m$, taken over the near-infinite population with all parameters fixed at their true values, is equal to

$$\beta_o^m \;=\; \mathbb{E}\,[\,Y_1 \,|\, Z=1\,] - \mathbb{E}\,[\,Y_0 \,|\, Z=1\,] \;=\; \beta_o^t.$$

Matching on the propensity score produces a consistent estimator of the population average causal treatment effect on the treated, β_o^t, which will be the same as the population average causal treatment effect, β_o, only when the individual-level treatment effect is independent of the propensity score. As discussed previously, whether we are more interested in estimating β_o or β_o^t will depend on the question we wish to answer.

If we were interested in estimating β_o but wished to use matching methods, we could apply more complex matching methods that match with replacement for both treated and control subjects and produce consistent estimators of β_o [1].

Inexact matching on the propensity score can introduce bias into the estimator $\hat{\beta}^m$. Different metrics on which to match [86, 114] or better matching algorithms [83] have been proposed to reduce this bias. Rubin and Thomas [93] found that regression adjustment on a sample matched on the propensity score was superior to either regression adjustment or propensity score matching alone, in terms of bias.

Estimators obtained from matched analyses also tend to have large variances, since the information contained in all unmatched subjects is discarded. This becomes a problem when a substantial proportion of the sample cannot be matched, which occurs frequently in epidemiological applications [102]. Solutions to this problem include matching with a variable number of controls [70, 97] and more complex forms of matching that use all the subjects [27, 36].

A couple of studies have considered the issue of making inferences from a propensity score matched treatment effect estimator. Theoretical properties of an estimator matched on the propensity score, modified to estimate the population average causal treatment effect, were studied by Abadie and Imbens [1]. They showed that their estimator is consistent and asymptotically normal, and derived an estimator for the variance conditional on the observed covariates and treatment status. Hill and Reiter [38] investigated methods of constructing confidence intervals for the matched estimator $\hat{\beta}^m$ using simulation studies, and found that bootstrap procedures were generally the most reliable.

2.2.4 Covariate adjustment including the propensity score

Covariate adjustment including the propensity score refers to the method of fitting a regression model for the outcome, which is allowed to depend on treatment status and propensity score, where usually the relationship between the outcome and the propensity score is assumed to be linear. Maximum likelihood regression models are a common method of adjustment for confounding in epidemiological studies, and have certain attractive properties. In particular, if the fitted regression model is correctly specified, the treatment effect estimator will be asymptotically unbiased [17, p.304]. Furthermore, as the sample size gets large, the variance of the treatment effect estimator will approach the Cramer-Rao lower bound. These properties will hold when using the covariate adjustment method. So, although this method does not appeal to the randomisation argument given in Section 2.1, if the fitted regression model is correct then the resulting treatment effect estimator will be asymptotically unbiased and will have the smallest possible variance. Rosenbaum and Rubin showed that when the propensity score is a monotone function of the linear discriminant, regression on the observed covariates is equivalent to regressing on the linear discriminant only — a function of the propensity score [84].

The covariate-adjusted treatment effect estimator

If the true relationship between the outcome and the propensity score is linear, then the treatment effect can be estimated by fitting a model where outcome depends linearly on the propensity score and treatment status. Denoting the unknown regression coefficients by ζ_0, ζ_1, and β_o^c, for the constant, the effect of the propensity score, and

the treatment effect, the model is

$$\mathbb{E}[Y] = \zeta_0 + \zeta_1 p(\mathbf{X}) + \beta_o^c Z. \tag{2.6}$$

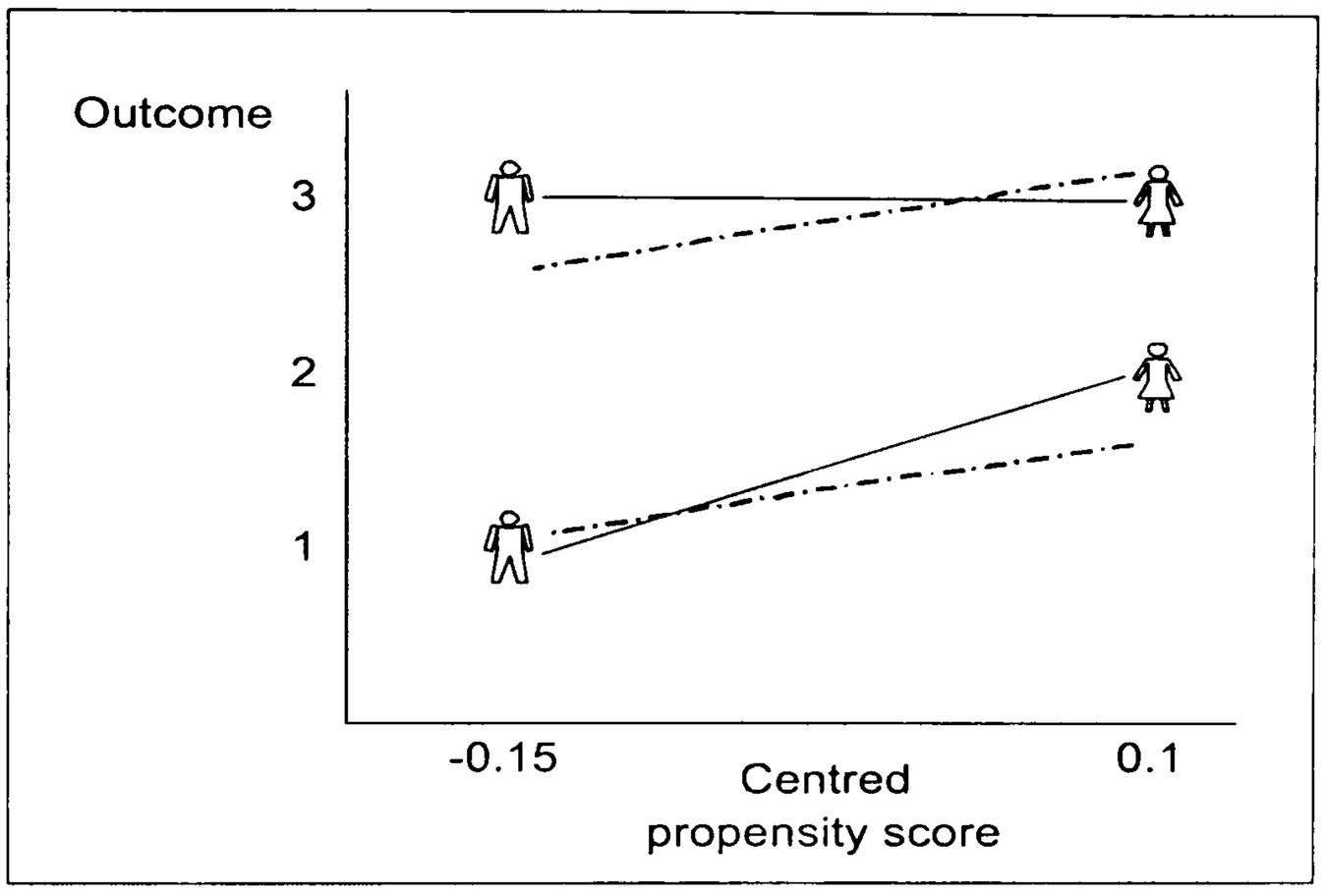
Fitting the regression model estimates β_o^c by $\hat{\beta}^c$ which is the covariate-adjusted treatment effect estimator.

Application to the example dataset

We have already estimated the propensity score for the artificial example given in Figure 2.1, finding that the girls have an estimated propensity score of 1/2 and the boys have an estimated propensity score of 1/4. We fit model (2.6) replacing the propensity score by the centred propensity score. This results in the same treatment effect estimate as model (2.6) but we later fit a more complex model which is more easily interpretable using a centred propensity score. To calculate the centred propensity score we subtract the sample mean of the estimated propensity score from each child's propensity score, giving boys a centred propensity score of -0.15 and girls a centred propensity score of 0.1. Using least-squares to fit model (2.6) produces the two dotted regression lines shown in Figure 2.4. This model assumes that the treatment effect is constant for all values of the propensity score — in other words, the same for both boys and girls — which is not true in this example. The model estimates that the effect of attending the homework club on educational achievement is $\hat{\beta}^c = 1.33$. This is equal to neither the population average causal treatment effect nor the population average causal treatment effect on the treated. Modifying model (2.6) to allow the treatment effect to vary with the propensity score results in the fitted regression lines shown in bold in Figure 2.4. This more complex model estimates that the effect of attending the homework club on educational achievement is $\hat{\beta}^c = 1.4$, the true population average causal treatment effect.

Since model (2.6) will only produce an unbiased estimator of β_o when the assumptions of a linear relationship between propensity score and outcome and a constant treatment effect are valid, we must consider carefully what these assumptions imply. In this artificial example, a higher propensity score is linked with a lower treatment effect. It would not be surprising to find, in an observational epidemiological study, that the physician tends to recommend a particular treatment only to those patients he feels will benefit from the treatment. If this were the case, the treatment effect would increase with the propensity score, a violation of the model assumptions in

Figure 2.4: *The covariate-adjusted analysis of the dataset shown in Figure 2.1, when the fitted regression line from model* (2.6) *is shown in dotted lines, and the fitted regression line from model* (2.6) *with an added interaction of propensity score and treatment is shown in bold.*



(2.6). Furthermore, in situations where the treatment effect varies with the propensity score, there is no clinical reason why it should do so linearly across the distribution of the propensity score, which is implied by the second, more complex model which was fitted in the example above.

When there is an interaction between the treatment effect and the propensity score the population average causal treatment effect, β_o and the population average causal treatment effect on the treated, β_o^t, will be different. When this occurs the methods of stratification on the propensity score and weighting by the inverse of the propensity score will both, without further modification, produce consistent estimators of β_o. Matching on the propensity score, again without further modification, will produce a consistent estimator of β_o^t. Only the covariate-adjusted method needs to be adapted to produce a consistent estimator of treatment effect.

The bias and variance of the covariate-adjusted estimator

If the linear model specified by (2.6) is correct, the error is independent of both treatment status and the propensity score, and the treatment effect is constant across the propensity score, then the asymptotic expectation of this estimator, taken over

our near-infinite population with all parameters fixed at their true values, is

$$\beta_o^c = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \beta_o.$$

The use of propensity scores was motivated in Section 2.1 by a randomisation argument showing that if all confounders were observed adjustment for the propensity score would produce an unbiased treatment effect estimator. This pseudo-randomisation property of propensity score methods is often seen as their primary advantage [93]. The covariate adjustment method, however, in making the added assumptions implicit in fitting model (2.6), loses this advantage [4].

If, rather than the linear propensity score term in model (2.6), we included a categorical variable indicating the quintiles of the propensity score, then the resulting estimator would approximate the stratified estimator of treatment effect obtained by stratifying at the quintiles of the propensity score, with added assumptions about independence of the error term and the constancy of the treatment effect. Therefore, it could be argued that a procedure which compared a regression model with outcome depending linearly on treatment status and a categorical propensity score, to model (2.6) and found that the latter fitted better, may result in a gain in precision from using a continuous covariate as opposed to a categorical variable, with little loss in bias in comparison with the stratified treatment effect estimator.

When the outcome surfaces are parallel — the treatment effect is constant across the propensity score — covariate adjustment including the propensity score has been found to reduce the bias of the treatment effect estimator [18]. Rosenbaum and Rubin [84] suggest allowing the treatment effect to vary with the propensity score in model (2.6) as we did in the artificial example. Important predictive covariates can also be added to model (2.6) in order to decrease the variance of the treatment effect estimator and to attempt to reduce bias [18].

2.2.5 *Weighting by the inverse of the propensity score*

The final propensity score method that we consider takes a different approach and does not directly use the randomisation argument outlined in Section 2.1. The outcomes of treated subjects are weighted by the inverse of the propensity score, $p(\mathbf{X})$, and the outcomes of untreated subjects are weighted by the inverse of $(1 - p(\mathbf{X}))$.

The resulting estimator is one of a larger class of estimators called inverse-probability-weighted estimators. The theoretical properties of these estimators are discussed extensively by Robins, Rotnitzky and Zhao [80] in the context of missing data. These methods can be directly applied to propensity scores by viewing whichever of the pair (Y_0, Y_1) is not observed as a missing observation. The general idea in inverse weighting is to create two 'potential' samples, that are intended to represent: (i) the sample we would have observed if everyone was treated, and (ii) the sample we would have observed if no-one was treated.

We reconstruct the 'potential' treated sample as follows. A subject with an estimated propensity score of 20% has a one-in-five chance of receiving the treatment. Therefore, for each treated sampled subject with an estimated propensity score of 20%, we assume that four others exist who were not treated, so we create four replica subjects, assigning these replicas the outcome of the initial treated sampled subject. Repeating this process for each value of the estimated propensity score, we reconstruct a potential treated sample that has the same size — or, if the estimated propensity score is continuous, approximately the same size — as the initial sample. Mathematically, this procedure is equivalent to weighting the outcome of each treated person by $1/p(\mathbf{X})$.

The same process is then followed to create a potential untreated sample. For each four untreated sampled subjects with an estimated propensity score of 20%, we assume that one subject exists who was treated. Hence we create a single replica for each four such untreated subjects, assigning the replica the mean outcome of the four subjects he replicated. Repeating this process for each value of the estimated propensity score, we reconstruct a potential untreated sample, using only the untreated subjects in the sample and the estimated propensity score. Mathematically, this procedure is equivalent to weighting the outcome of each untreated person by $1/(1-p(\mathbf{X}))$.

The mean outcomes of the potential treated and untreated samples are unbiased estimators of the mean outcome of the whole population if everyone were treated or everyone were untreated, respectively. Therefore, the difference in the mean outcomes in these two potential samples is an unbiased estimator of the population average causal treatment effect.

The inversely-weighted treatment effect estimator

The inversely-weighted treatment effect estimator produced by the process described above is

$$\hat{\beta}^w = \frac{\sum_{i=1}^{n} \frac{Y_i Z_i}{p(\mathbf{X}_i)}}{\sum_{i=1}^{n} \frac{Z_i}{p(\mathbf{X}_i)}} - \frac{\sum_{i=1}^{n} \frac{Y_i (1-Z_i)}{(1-p(\mathbf{X}_i))}}{\sum_{i=1}^{n} \frac{(1-Z_i)}{(1-p(\mathbf{X}_i))}}.$$

The two sums in the denominator merely ensure that the weights for treated and untreated subjects both sum to one.

Application to the example dataset

We previously estimated the propensity score for the example dataset, finding that girls had an estimated propensity score of 1/2 and boys had an estimated propensity score of 1/4. We first construct a potential treated sample that is intended to represent the sample we would see if everyone were treated. Since each treated girl has an estimated probability of 1/2 of receiving treatment, we assume that for each of these treated girls in the sample, there is one untreated girl in the sample with the same propensity score. Therefore, in our potential treated sample, a single replica of each of these treated girls is created, and allocated the same outcome as the treated girl, which in this case is 3. Similarly, the two treated boys have a one-in-four chance of receiving treatment. Therefore, we create three replicas of each treated boy. The potential treated sample is shown in the top half of Figure 2.5. This potential treated sample contains both the eight sampled children who were treated, depicted by unshaded figures, and the twelve replicas of these treated subjects, depicted by shaded figures. Similarly, the potential untreated sample was created from the untreated sampled subjects, using the estimated propensity score.

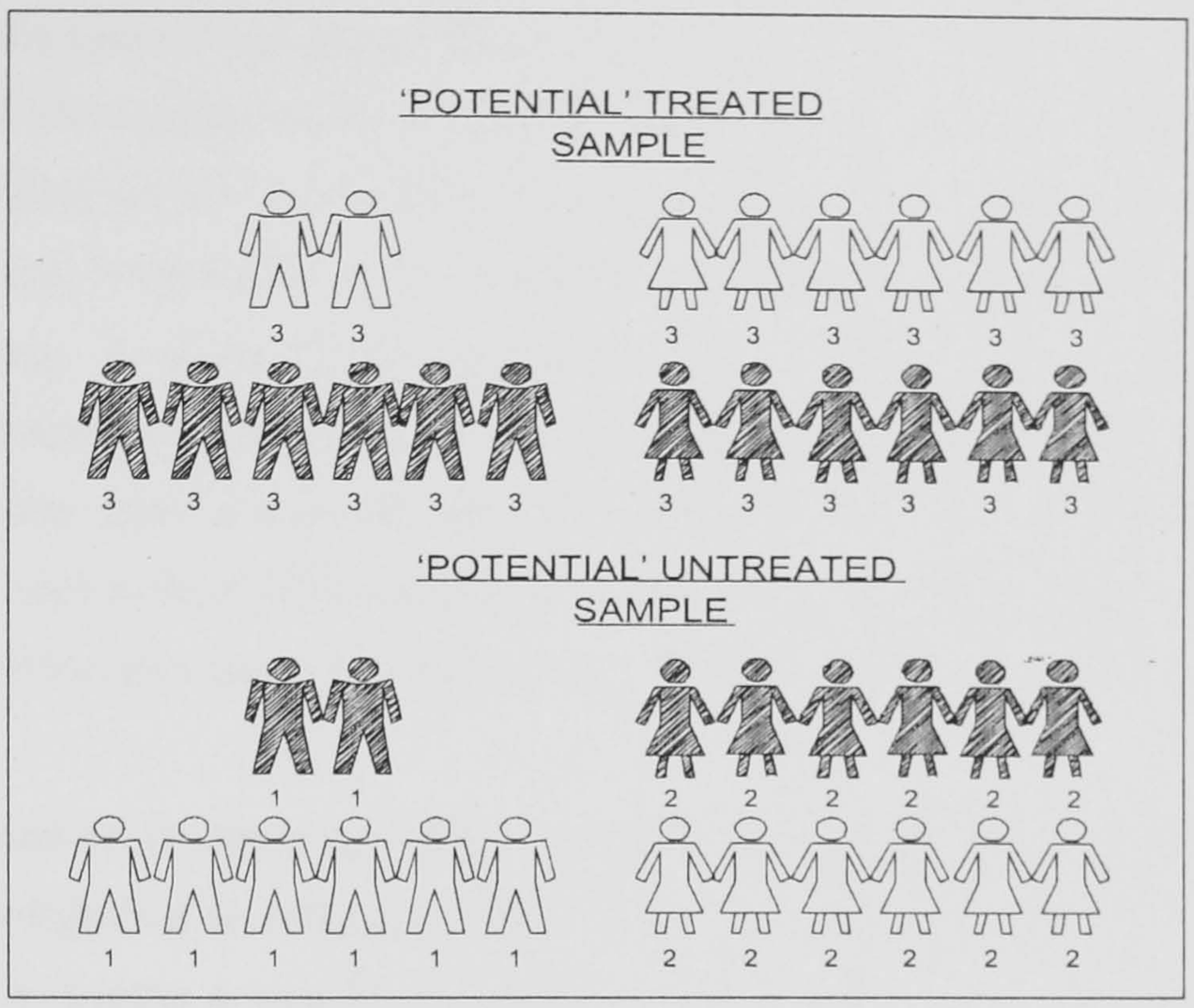
The mean outcome in the potential treated sample is 3. The mean outcome in the potential untreated sample is $(12 \times 2 + 8 \times 1)/20 = 1.6$. Therefore, the inverse-weighting method gives an treatment effect estimator of $\hat{\beta}^w = 1.4$, which is equal to the population average causal treatment effect β_o.

The bias and variance of the inversely-weighted estimator

The asymptotic expectation of the estimator $\hat{\beta}^w$, taken over the near-infinite population with all parameters fixed at their true values, is

$$\beta_o^w = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \beta_o.$$

Figure 2.5: *The weighted analysis. Two 'potential' samples are created from the initial sample. Each 'potential' sample contains the initial treated or untreated subjects (unshaded) and replicated subjects (shaded) whose addition is intended to create two samples which both have the same covariate structure as the whole sample.*



The variance of the estimator β^w, when the propensity score is known or consistently estimated, has been calculated using M-estimation methods[4] by Lunceford and Davidian [59]. For treated subjects who have a propensity score close to zero, or untreated subjects who have a propensity score close to one, the weight can be extremely large, producing an estimator with extremely large variance.

More complex inversely-weighted estimators have been proposed. Both Hirano and Lunceford and Davidian give estimators that have the smallest possible variance of all such semi-parametric inversely-weighted-probability estimators. Hirano's estimator has the advantage of only requiring modelling of the propensity score [40]. Lunceford and Davidian's version requires both the outcome and the propensity score to be modelled [59]. However, their estimator is doubly robust: if either the model for the propensity score or the model for the conditional mean of the outcome, given treatment status and covariates, is wrong then the estimator will still be consistent, although it may then have a larger variance.

[4]M-estimation methods, also called estimating equation methods, can be used to make asymptotic inferences from an estimator without specifying the full probability distribution of the data. For more details see Section 3.1.2 and references therein.

Doubly robust methods are particularly attractive in that it is only necessary to specify one of two models correctly, giving the analyst an extra layer of protection against bias. Although doubly robust propensity score methods are a promising method of analysis we do not pursue them any further here. This is because much is already known about the theoretical properties of general doubly robust estimators and others are researching their application within a propensity score context [59]. Furthermore, as we will see (Section 2.3.3) epidemiologists appear to be reluctant to use the simpler inverse-weighting propensity score methods, perhaps due to unfamiliarity or a lack of understanding. It seems likely, therefore, that it will take time for doubly robust methods to become widely used in epidemiological studies. Since one of the aims here is to provide practical guidance for epidemiologists, we have chosen to investigate improved methods of inference for a currently used method rather than studying a possibly superior but more complex and infrequently used method.

If we were interested in the population average causal treatment effect on the treated, we could use weighting to estimate this by assigning treated subjects a weight of one and untreated subjects a weight of $p(\mathbf{X})/(1-p(\mathbf{X}))$ [55].

2.3 Comparison of propensity score methods

We have so far discussed the theoretical justification for propensity score methodology and described the four main propensity score methods. An important question now is: which of these methods should be used in applications? Are there situations in which one method will be 'better' than the others? Whilst a comprehensive answer to these questions is beyond the scope of this thesis, we attempt to show how the four methods are inter-related and to use this knowledge to explain various well-known features of the four estimators. We then consider the implications of estimating the propensity score for each method. We review analyses that have compared the various methods, and end with a few remarks about the comparative uptake of the four propensity score methods in epidemiology.

2.3.1 Links between the propensity score methods

In the previous section, we applied each of the four propensity score methods to an artificial example dataset. Inspection of the way in which these estimators was

constructed should convince us that when the individual-level treatment effect is the same for all values of the propensity score, there is no error in the outcome, and the propensity score is discrete, then stratification and matching on the propensity score, and weighting by the inverse of the propensity score all produce exactly the same estimator. Covariate adjustment including the propensity score, when the fitted regression model is correctly specified, will also give the same estimator. So in a very basic situation, the four methods are essentially identical. We now consider what happens when each of the three simplifying conditions above is relaxed.

Non-uniform treatment effect

If the treatment effect is not the same for all values of the propensity score, then the two estimands β_o and β_o^t will be different. Stratification on the propensity score, weighting by the inverse of the propensity score and covariate adjustment including the propensity score all estimate β_o, whilst matching on the propensity score estimates β_o^t. In fact, each of these methods can be adapted to estimate either β_o or β_o^t, but these more complex versions are not frequently used in applications.

Error in the outcome

In practice, treatment status and covariates do not uniquely determine the outcome. This variation in outcome is due to, for example, random error, measurement error and unobserved non-confounding variables. We still assume that there are no unobserved confounders but now allow the outcome to contain random error. For all propensity score methods this leads to error in the estimator. However, the implications of this error are more important for the matching and inverse-weighting approaches. Typically, when matching on the propensity score a large number of subjects are unmatched and therefore discarded. When there is error in the outcome, any information lost in discarding unmatched subjects will increase the variance of the treatment effect estimator. It is possible that this increase may sometimes be substantial.

When weighting by the inverse of the propensity score, error in the outcome means that when some treated subjects have small propensity scores, or some untreated subjects have large propensity scores, then the treatment effect estimator can have extremely large variance. In order to understand why, let us return to the artificial dataset shown in Figure 2.1. Suppose that the outcome of one of the two treated boys was wrongly measured as 5, rather than 3. If we were to use this new measurement

when reconstructing the potential treated sample, the three replicas of this boy would also be assigned an outcome of 5. The potential untreated sample would remain the same, resulting in a treatment effect estimate of $\hat{\beta}^w = 1.8$, larger than the true value, $\beta_o = 1.4$. In this case, a small measurement error created a small bias in the estimator. In more extreme cases the effect of a small error is much more striking. Suppose that the treated boy in question had a propensity score of 1/1000 rather than 1/4. We would then create 999 replicas of him, rather than 3. The error in his outcome would be replicated 999 times, possibly resulting in a substantial bias in the treatment effect estimate. Thus error in the outcome will have the largest effect for subjects who are replicated many times. This will happen when either the subject is treated and has a very low propensity score, or the subject is untreated and has a very high propensity score.

Continuous propensity score

When some covariates are continuous it is unfeasible to estimate the propensity score at each different set of covariate values by the fraction of subjects with those covariate values who are treated, as in the artificial example. Typically, in epidemiological applications, a logistic regression model is used to estimate the propensity score [111].

A small amount of bias may be introduced into the matching estimator due to inexact matching. In the same way, bias can introduced into the stratified estimator due to the strata being non-homogenous in the propensity score. This is often referred to as residual confounding [59].

The effect of a continuous propensity score on the weighting method is more complicated. In the discrete case the propensity score is estimated by the fraction of treated subjects at each combination of covariate values. Then, for example, for each treated subject in the sample with an estimated propensity score of 1/3, there will be two untreated subjects in the sample with an estimated propensity score of 1/3. Therefore, when we reconstruct the two potential samples, each treated subject with this propensity score is replicated twice, and each two untreated subjects with this propensity score are replicated once. This ensures that each of the two potential samples being compared have exactly the same number of subjects with that estimated propensity score. Therefore, we are comparing two 'identical' populations, in terms of the distribution of the propensity score. Now suppose that the propensity score is continuous. In this case, a treated subject with a propensity score of 1/3 is likely to be

the only subject in the sample with exactly that propensity score. We would replicate this subject twice, resulting in the potential treated sample containing three subjects with an estimated propensity score of 1/3. Conversely, since there are no untreated subjects with that propensity score, the potential untreated sample would contain no subjects with an estimated propensity score of 1/3. Because of this, the two potential samples are no longer identical, in terms of their propensity score distributions, although in large samples they should be similar. This problem is particularly relevant when the tail end of the propensity score distribution is dominated by one treatment group. For example, if no treated subjects have low propensity score values and many untreated subjects do, then the two reconstructed potential samples may end up being very dissimilar, producing a biased treatment effect estimator. In order to avoid such problems, a common support condition can be imposed which ensures that the range of propensity score values in the two treatment groups is the same [36]. This is a sensible criterion to impose when using any propensity score method, but the implications of not doing so will be most damaging when weighting by the inverse of the propensity score.

Links with stratification

We have seen that in very simple, unrealistic applications, all four propensity score methods will produce the same estimator. In more realistic settings, the different approaches taken by the four methods produce different estimators. In order to further understand the comparative benefits of each method, we now draw links between each method and the method of stratification on the propensity score. Stratification is taken as the 'baseline' method since, in our opinion, it is the method which most closely reflects the randomisation argument that we used to motivate the propensity score approach.

- Matching on the propensity score can be viewed as an extremely fine stratification on the propensity score. Discarding all unmatched subjects loses information, and hence increases the variance of the treatment effect estimator, and also changes the composition of the study population, often leading to a change in the estimand.

- Stratification on the propensity score can be seen as a weighted estimator. Suppose we were unsure how to model the propensity score. We could use a logistic regression model to get an 'initial guess' at the propensity score. Although

we expect this model to be wrong, we might hope that it is good enough to be able to group subjects into rough classes containing subjects with similar propensity scores. The propensity score could then be estimated separately for subjects in each of these classes by the fraction of subjects in that class who are treated. We would then calculate the weighted treatment effect estimator $\hat{\beta}^w$ using this new propensity score. This is exactly equivalent to estimating the treatment effect using stratification on the propensity score. Viewing the stratified estimator in this way, we would expect that if the propensity score is modelled correctly, the weighted estimator would be less biased than the stratified estimator, but we might expect the stratified estimator to be more robust to mis-specifications of the propensity score model, and to be less sensitive to observations with particularly high or low propensity scores.

- Covariate adjustment including the propensity score applies extra mathematical constraints to the data, in exchange for added power. As has been mentioned, if the treatment effect is the same for all values of the propensity score, fitting a model where the outcome depends on treatment and a categorical variable indicating the quintiles of the propensity score is an approximation to stratification on the propensity score. Including the propensity score as a linear covariate adds yet more assumptions, and unless the model is correct, this risks introducing bias.

2.3.2 *Implications of estimating the propensity score*

Having investigated the links between the four propensity score treatment effect estimators, we now consider the implications of estimating the propensity score. We begin by revisiting the artificial example introduced earlier in this chapter (see Section 2.2.1). Suppose the minister for education subsequently found out that the children in her dataset had not been given the choice of attending the homework club. The headmaster had decided to make the new club compulsory for a random selection of 40% of his pupils, in order to give it a trial-run before making the policy school-wide. In other words, the 'true' propensity score — the probability of attending the homework club — is 0.4 for both boys and girls and the fact that more girls than boys were selected is merely due to chance. Armed with this new information, we re-analyse the dataset with the four propensity score methods using the 'true' propensity score.

With only one propensity-score value, stratification on the propensity score estimates the treatment effect by the difference in mean outcomes between the treated and untreated subjects in the whole sample. This gives $\hat{\beta}^s = 1.5$, higher than the true population average causal treatment effect, $\beta_o = 1.4$, due to the lower baseline outcome of boys, and the over-representation of girls in the treatment group. Since two subjects with the same propensity score can now have different covariate values, the estimate produced by matching depends on the selection of the matched pairs. If we were really unlucky, we might match each of the treated girls to an untreated boy, and each of the treated boys to an untreated girl. This would estimate the treatment effect by $\hat{\beta}^m = 1.75$, much higher than either β_o or β_o^t. Covariate adjustment including the propensity score, since there is only one value of the propensity score, is now equivalent to estimating the treatment effect by the difference in mean outcomes between the treated and untreated subjects, and so $\hat{\beta}^c = 1.5$. Weighting by the inverse of the propensity score would produce a 'potential' treated sample consisting of 15 girls and 5 boys, all with an outcome of 3, and a 'potential' untreated sample consisting of 10 girls, with an outcome of 2, and 10 boys, with an outcome of 1. This also estimates the treatment effect by $\hat{\beta}^w = 1.5$.

This example illustrates a phenomenon that has often been observed [84, 85] — that adjusting for the estimated propensity score often performs better than adjusting for the true propensity score. In an infinite sample, for any propensity score p, a fraction of *exactly* p of the subjects with that propensity score would be treated. Therefore, adjusting for the 'true' propensity score would give an unbiased treatment effect estimator. In finite samples, however, imbalances may arise due to chance. In this case adjusting for the 'true' propensity score adjusts only for systematic imbalances whereas using the estimated propensity score also adjusts for imbalances due to 'bad luck' [49]. There are two ways in which we can adjust for bad luck as well as systematic differences. The first is when we use the correct model for the propensity score, but estimate the coefficients. The second is called overfitting the propensity score model — including covariates related to outcome that are not truly related to treatment but happen to be unbalanced in our sample. Rosenbaum shows that a simple stratified estimator obtained by stratifying on the 'true' propensity score is unbiased but not conditionally unbiased given the observed data, whereas stratifying on an over-fitted propensity score produces a conditionally unbiased estimator [82], which is exactly what we have seen from the two analyses of the artificial dataset. We expect that removing a conditional bias will reduce the variance of the estimator. In line with this

intuition, both simulation studies [32] and practical applications [84, 85] suggest that estimation of the propensity score can decrease the variance of the treatment effect estimator. Theoretical results show that the large-sample variance of a treatment effect estimator obtained using weighting by the inverse of the propensity score is reduced by both estimating and overfitting the propensity score [59]. Although no analogous theoretical results have yet been proved for the methods of stratification or matching on the propensity score, there is some empirical evidence that estimation and overfitting of the propensity score have similar effects in these situations [59]. The effects on the covariate-adjusted estimator have not been investigated.

If overfitting the propensity score model leads to a reduction in bias, then missing out a covariate related to outcome that is unbalanced in our sample — a confounder — will lead to an increase in bias. Simulation results have demonstrated that large biases can occur when a confounder is omitted from the propensity score model when stratifying or matching on the propensity score [24, 115]. In fact, the above discussion shows that this will be true of all propensity score methods. This result and the results contained in the previous paragraph may encourage us to add every observed covariate into the logistic regression model for the propensity score. It has, however, been pointed out that the propensity score must be consistently estimated in order for the treatment effect estimator to be consistent, and therefore, there is a limit to the number of variables that can be entered into the propensity score model [50, 79].

We now move on to consider the effect of wrongly specifying the form of the propensity score model. Simulation studies suggest that incorrectly specifying the form of the model fitted for the propensity score, for example applying a logistic model when the true propensity score is linear, introduces little bias when stratifying or matching on the propensity score [24, 115]. Since weighting by the inverse of the propensity score attaches more importance to observations at the tail-end of the propensity score distributions, and fitting a linear rather than a logistic regression model would probably only differ substantially at the tail-ends of the distribution, we may expect the effect of fitting the wrong type of model to affect the weighted estimator more than either the stratified or matched estimators. There is little evidence concerning wrongly specifying non-linearities or omitting important interactions.

Propensity score diagnostics

Weitzen *et. al.* [111] review epidemiological applications of propensity score methods and find no consensus about which model diagnostics are appropriate. Both

stratification and matching on the propensity score aim to create groups of subjects within which the observed covariate distributions are balanced. As long as this balance is achieved, it does not matter how badly the propensity score was estimated. In particular, if our estimated propensity score is wrong, but is a monotone function of the true propensity score, then the stratified and matched estimators will still be consistent. In view of this, the only diagnostics needed when stratifying or matching on the propensity score are ones which assess the balance achieved within the strata or matched pairs [92]. Weighting by the inverse of the propensity score attempts to use the estimated propensity score to construct two potential samples, both of which have covariate distributions representing the whole sample. Since the estimated values of the propensity score are used to create these potential samples, we might expect that it is more important that the model should be correct than with the stratified or matched estimators, and so it may be necessary to use model diagnostics that assess the accuracy of the estimated coefficients in the propensity score model. Similarly, standard model diagnostics may be necessary when using covariate adjustment including the propensity score, since in this case, the precision of the estimated propensity score is likely to impact on the treatment effect estimator.

2.3.3 Review of empirical comparisons of propensity score methods

We now review studies that have compared the various propensity score methods either with each other or with standard regression models. Most of these studies have applied propensity score methods to the estimation of odds ratios or hazard ratios, rather than differences in means, as described previously in this chapter. We will return later to the issue of discrete outcome data.

Two papers have reviewed the epidemiological literature, looking for studies that have applied both standard regression models and at least one propensity score method in order to estimate a causal effect. The first found 43 studies containing 54 estimated odds ratios or hazard ratios [95]. They found that 8/54 of these had more than a 25% difference in the estimated effect (on the log scale) produced by standard regression and at least one of the propensity score methods. However, in none of these were the two point estimates both significant and on either side of unity, and in most cases the regression estimate was significant but the propensity score estimate was not. It appears that estimates from a regression model, as expected, are more likely to

produce significant results. However, these results are not sufficient to say whether this is because the regression model is appropriate, or whether regression models are giving us an unjustified sense of security. The second review paper found 69 estimated effects, 9/69 of which had more than a 20% difference in the estimated effect when compared with any of the propensity score estimates given [102]. In a further 4/69 of the studies, the covariate-adjusted estimate was the only propensity score method which differed substantially from the regression estimate, and in 1/69 the matched estimate was the only propensity score estimate that differed from the regression estimate. The difference between the matched estimate and the regression estimate in this final study may be attributable to the difference in the two estimands that are analogous to β_o and β_o^t on the odds ratio scale. This cannot be said for the covariate-adjusted estimates.

Several analyses have been conducted using various propensity score methods and regression models specifically for the purposes of comparison. Sturmer [103] investigated the effect of nonsteroidal anti-inflammatory drug use on one-year all-cause mortality, using several propensity score methods applied to Cox survival models. Stratification and matching on the propensity score, covariate adjustment including the propensity score, weighting by the inverse of the propensity score and a standard Cox regression model all produced effect estimates comparable with the randomised evidence of no effect. A second analysis studied the effect of statin therapy on all-cause mortality following acute myocardial infarction [4]. Various propensity score methods and standard regression methods were applied to the observational dataset, all of which gave fairly comparable point estimates, but the estimate from the regression model had the lowest variance. The estimate produced by matching on the propensity score gave an identical point estimate to a meta-analysis of trials, perhaps because the composition within the matched sample was younger and healthier — and therefore more similar to participants in the randomised trials. A final paper considered the effect of tissue plasminogen activator on death among ischemic stroke patients, using data from a German stroke registry, an example in which there was strong evidence of the treatment effect varying with the propensity score [55]. Using the whole dataset, matching on the propensity score, and weighting by the inverse of the propensity score, modified to estimate the population average treatment effect on the treated, both produced estimates comparable with the randomised evidence. Stratification on the propensity score, covariate adjustment including the propensity score, and standard regression models all appeared to overestimate the treatment

effect, whereas the unmodified method of weighting by the inverse of the propensity score appeared to overestimate the treatment effect to a factor of about 10, due to the presence of many very small propensity score values. Since the treated subjects in the observational dataset appeared to more closely resemble those who participated in the trials, it is perhaps unsurprising that the two methods that estimate β_o^t produced estimates closer to the randomised evidence. Restricting the sample to those subjects with a non-negligible propensity score — greater than 0.05 — resulted in all treatment effect estimates being fairly comparable with the randomised evidence.

Theoretical results that compared weighting by the inverse of the propensity score with stratification on the propensity score, with additional within-stratum regression, found that the stratified estimator outperformed the weighted estimator in some situations but that the inverse-weighting method can be made 'doubly-robust' adding a layer of protection for the analyst, although what happens when both models are incorrect is unclear [59].

Overall, it is not clear whether any of the four methods of using the propensity score is superior to the others. Neither is it clear whether propensity scores perform better or worse than standard regression models in practice.

The uptake of the four propensity score methods

Weitzen *et al.* [111] carried out a systematic literature review of all studies investigating a health or medical related question that were published in 2001, identifying 47 studies using propensity score methods. Of these studies, 24 used covariate adjustment including the propensity score, 13 used stratification on the propensity score, and 8 used matching on the propensity score. The remaining 2 studies did not report the particular propensity score method used. None of the studies reported using weighting by the inverse of the propensity score.

2.4 Extensions of propensity score methods

It is beyond the scope of this thesis to discuss all avenues of research concerning propensity scores. We briefly draw attention to four major extensions of propensity score methods: the application of propensity score methods to datasets with multiple treatments, missing data methods, the extension of propensity score methods to more

complex study designs, and the application of propensity score methods from a non-frequentist perspective.

The propensity score, as defined above, only applies to situations where there is one treatment of interest and one control treatment. Joffe and Rosenbaum [49] first studied the application of propensity score methods to situations with more than one level of treatment. They proposed a multiple propensity score $p(t, \mathbf{X}) = \mathbb{P}(Z = t \mid \mathbf{X})$ where the treatment status indicator Z takes multiple values. The distribution of Z must depend on the covariates only through this multiple propensity score, so that $\mathbb{P}(Z = t \mid \mathbf{X}) = \mathbb{P}(Z = t \mid p(t, \mathbf{X}))$. This will happen, for example, if treatment allocation follows a proportional odds model. If such a multiple propensity score exists, then treatment is strongly ignorable given the multiple propensity score, since at any value of this multiple propensity score all potential outcomes are independent of the treatment status indicator. Imbens [46], however, notes that there is often no multiple propensity score that fulfils the criterion above but that weaker conditions can be imposed that are less likely to be violated, in which case propensity score methods can still be applied.

Another important issue in observational epidemiology is that of missing data, since it results in a loss of precision and may lead to biased treatment effect estimates [58]. Rosenbaum and Rubin [85] suggested using the 'generalized propensity score' which has the same balancing properties as the usual propensity score. The generalized score is defined as $\mathbb{P}(Z = 1 \mid \mathbf{X}_{obs}, R)$ where $\mathbf{X}_{obs}$ denotes the observed covariate information and R is a missing data indicator. For discrete covariates, this is equivalent to adding an extra category of 'missing' to each partially-observed covariate. This approach, however, can be impractical if there are too many missing data patterns. More complex versions of this approach have been suggested [19].

We have so far concentrated on the analysis of cohort studies. Propensity score methods, however, have been used in other contexts. For example, they have been applied to multilevel data [42], survival data [104], survey data [113], repeated cross-sectional data [36], and case-control studies [49]. Weighting by the inverse of the propensity score has been used in the context of marginal structural models in order to estimate the effect of time-varying treatments [78] and similar ideas have also been used to estimate treatment effects with censored time-lagged data [3].

We end with a note on non-frequentist perspectives. It is possible to use propensity score methodology to create permutation tests of the null hypothesis of no treatment effect for any sampled subject, using the randomisation distribution induced by the treatment allocation mechanism [81]. These tests are comparable with Fisher's exact test and do not appeal to the idea of repeated sampling from a near-infinite population. Finally, there has recently been some interest in applying propensity score methods within a Bayesian framework [65].

2.5 Discussion

In this chapter we have used a randomisation argument to motivate the use of propensity scores to estimate causal treatment effects when confounding is present, demonstrating that when all confounders are observed, propensity scores can be used to create a pseudo-randomised situation, allowing unbiased estimation of causal effects. Four main propensity score methods were described in detail: stratification on the propensity score, matching on the propensity score, covariate adjustment including the propensity score, and weighting by the inverse of the propensity score.

A commonly used method of analysing observational data in epidemiology is a maximum likelihood regression model of the outcome. The question we must now ask is whether there is any benefit in using propensity score methods, rather than standard regression models. Robins and Mark [79] point out that using the correct regression model for the outcome will always be more efficient than modelling the propensity score. Regression, however, has some important limitations. Firstly, regression models are highly dependent on the form of the fitted model. For example, Rubin [89] presents simulation results that show that regression can produced biased estimates when the individual-level treatment effect is not constant. Specifying the correct model is particularly difficult when information in the data about the form of the outcome is scarce, as is the case when the outcome is a rare event [11]. A second problem with regression is due to non-comparability of treatment groups in observational studies. When this occurs, the problem will be immediately apparent from a propensity score analysis [91]. Conversely, with a regression approach, this non-comparability will not be evident and the resulting treatment effect estimator will be essentially based on extrapolation [59]. It has been suggested that some sort of summary of the propensity score distributions in the treated and untreated groups

should be provided in any analysis of observational data, whether or not propensity score methods are used for the primary analysis [47]. A final problem with standard regression models is the choice of covariates to include in the model. These can be selected on the basis of either their intrinsic interest, their ability to predict outcome or their ability to predict treatment. There is no clear method for choosing which criteria to use. It has been suggested that all modelling strategies contain implicit prior beliefs and hence all modelling strategies should be recognized as an approximation to a formal Bayesian analysis [77]. Propensity score methods offer an alternative approach to the problem of estimating causal treatment effects, which may overcome some of these problems associated with standard regression models.

We have seen that covariate adjustment including the propensity score is the most widely used propensity score method in epidemiological applications, despite little theoretical or clinical justification for its use. The assumptions implicit in this method are unlikely to hold in practice. We have seen that occasionally the covariate-adjusted estimate is dissimilar to both the standard regression estimate and the estimates obtained using the other propensity score methods. It seems likely, in these cases, that the covariate-adjusted estimate is the incorrect estimate. The other three propensity score methods discussed — stratification and matching on the propensity score, and weighting by the inverse of the propensity score — each have advantages and disadvantages. Matching is conceptually simple and a familiar method to epidemiologists. Simple matching strategies, however, may estimate a quantity different to the one required. Weighting by the inverse of the propensity score does not appear to be much used in epidemiological applications, perhaps due to unfamiliarity. In datasets with many extreme propensity score values, the variance of the weighted estimator may be extremely large. The final method, stratification on the propensity score, is perhaps the most basic of the methods. It is computationally easy and conceptually simple. Although the stratified estimator is biased, there exist simple remedies for this. We can create more strata, making the groups more comparable and thus reducing the bias. We can also fit regression models within each strata, containing the major predictive variables. This serves both to decrease bias and to allow us to investigate the way in which risk factors affect outcome, within the propensity score strata. We therefore feel that the stratified estimator offers a simple and flexible method of estimating causal treatment effects for epidemiological practitioners.

This motivates our thorough theoretical investigation, hitherto lacking, to establish precisely when stratified estimators are consistent and normally distributed. This throws up some unexpected conditions with important practical implications. We are then able to derive theoretical asymptotic variance formulæ for the stratified treatment effect estimator facilitating, for the first time, a general comparison with, amongst other methods, standard regression models. Chapter 3 begins these calculations by investigating the large-sample theoretical properties of the stratified treatment effect estimator.

Chapter 3

Theoretical properties of the stratified treatment effect estimator

3.1 Introduction

We discussed, in Chapter 2, the justification for the use of propensity scores to estimate causal treatment effects in the presence of confounding, and described four main propensity score methods. We now focus on one of these methods in particular: stratification on the propensity score. We derive theoretical properties for the treatment effect estimator obtained using stratification on the propensity score, and establish conditions under which this estimator is consistent, derive its asymptotic sampling distribution, and calculate its asymptotic variance.

We begin by reviewing the notation given in the previous chapter, and introducing some additional notation. We then briefly outline the theory of M-estimation, since this will be used later in the chapter to calculate asymptotic variances. Although the propensity score is invariably unknown, and therefore must be estimated from the data, the variance formula typically used in epidemiological applications does not take account of this estimation, and treats the propensity score as if it were measured without error [59]. Therefore, we begin with the unlikely assumption that the propensity score is a known function of the observed data, and derive theoretical properties for this simplified estimator. We then extend this to the more realistic situation where the propensity score is estimated using a logistic regression model, and derive theoretical properties for the resulting estimator. The asymptotic variances of the two estimators, denoted by $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$ for the situations where the propensity score is known and estimated, respectively, turn out to be distinct. We show how the two variances can be expressed in terms of four variance components, reflecting the four different sources of variation affecting the stratified treatment effect estimator.

Finally, we consider the practical implications of these four variance components for the estimation of the propensity score.

3.1.1 Further notation

We use a subscript 'o' to denote a population parameter, and a hat to indicate a parameter estimator. For example, the stratified treatment effect estimator — the estimator obtained by stratification on the propensity score — is denoted by $\hat{\beta}^s$, which is the sample estimator of the population parameter β_o^s, which we will call the population stratified treatment effect. The population average causal treatment effect, the population parameter of interest, is denoted by β_o.

As before, we assume that the outcome is continuous and is denoted by Y. At intervals in the argument, it is convenient to let Y_1 denote a particular subject's outcome had they received treatment and Y_0 denote their outcome had they not received treatment (see Section 2.1). The treatment, denoted by Z, is assumed to be binary, where $Z = 1$ indicates that the subject received treatment and $Z = 0$ indicates that the subject did not receive treatment. Each subject has a vector of observed covariates, denoted by $\mathbf{X} = (X_1, ..., X_m)$. A sample of data, $\{Y_i, Z_i, \mathbf{X}_i\}$ for $i = 1, ..., n$, is drawn independently from the population.

We denote the propensity score by $p(\mathbf{X}) = \mathbb{P}(Z = 1 \,|\, \mathbf{X})$, which is assumed to be a continuous function. We now introduce some new notation to describe the propensity score and the strata. The propensity score is assumed to be a linear function of the observed covariates on the logit scale, parameterized by a vector of parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)$. Since the propensity score depends on these — possibly unknown — parameters, we need to distinguish between the true and estimated propensity score. We write these as $p_o(\mathbf{X})$ and $\hat{p}(\mathbf{X})$, when the parameter $\boldsymbol{\alpha}$ is known and estimated, respectively. It will sometimes be useful to emphasize the dependence on $\boldsymbol{\alpha}$, in which case we write $p(\mathbf{X}\,;\,\boldsymbol{\alpha}_o)$ and $p(\mathbf{X}\,;\,\hat{\boldsymbol{\alpha}})$. The probability density function of this propensity score is denoted by $f_p(\,\cdot\,)$.

In order to define the stratified treatment effect estimator, we must choose how many strata to use and the fractions in which the population should be divided between the strata. The number of strata is denoted by K and the fractions of the population

contained in each stratum are denoted by $\mathbf{r} = (r_1, r_2, ..., r_K)^T$. In applications, the sample is typically split into five equal strata, so $K = 5$ and $r_1 = r_2 = .. = r_5 = 1/5$. As we choose these numbers, the fraction of the population in each population stratum will be the same as the fraction of the sample in each sample stratum.

The quantiles of the propensity score distribution that split the population into strata containing fractions $(r_1, ..., r_K)^T$ are denoted by $\mathbf{q_o} = (q_{1o}, q_{2o}, ..., q_{(K-1)o})^T$. The fractions of treated subjects in the population contained in each stratum are denoted by $\mathbf{d_o} = (d_{1o}, d_{2o}, ..., d_{Ko})^T$. The fractions of untreated subjects in each stratum are then $(r_1 - d_{1o}, ..., r_K - d_{Ko})^T$.

As in Chapter 2, we let $\mathbf{S} = (S_1, ..., S_K)$ be a set of stratum indicators. We now add hats when we refer to the sample strata, in order to emphasize that these are estimators. So for subject i, $\hat{S}_{si}$ is equal to one if and only if that subject is in the s^{th} sample stratum, or in other words, $\hat{S}_{si}$ is equal to one if and only if $\hat{q}_{(s-1)} \leq p(\mathbf{X}_i) < \hat{q}_s$, for subjects $i = 1, ..., n$, and strata $s = 1, ..., K$. Whether the propensity score is known or estimated, when referring to $\hat{S}_{si}$, should be clear from the context. To simplify this notation, we let $1_{[A]}$ refer to the indicator function of event A, taking the value 1 if event A occurs, and 0 otherwise. Then we can write $\hat{S}_{si} = 1_{[\hat{q}_{(s-1)} \leq p(\mathbf{X}_i) < \hat{q}_s]}$. In line with the subscript notation above, $S_{so} = 1_{[q_{(s-1)o} \leq p_o(\mathbf{X}) < q_{so}]}$ is an indicator for the s^{th} population stratum.

Vectors and matrices are written in bold. Dimensions of matrices are indicated by superscripts, so $\mathbf{A}^{K \times 2}$ denotes a matrix of dimension $K \times 2$. Positions within a matrix are indicated by subscripts on a bracketed matrix, so the $(1,1)^{th}$ component of matrix $\mathbf{A}$ is denoted by $(\mathbf{A})_{11}$. By contrast, subscripts on matrices without brackets refers to one of a series of connected matrices, so for example $\mathbf{a}_{11}$ refers to a whole matrix rather than indicating a position within the matrix $\mathbf{a}$. Superscripts of 'T' on a vector or matrix denote its transpose and superscripts of '-T' denote the transpose of the inverse. Variance, covariance, expectation and probability are denoted by $\mathbb{V}[\cdot]$, $\mathrm{Cov}[\cdot]$, $\mathbb{E}[\cdot]$ and $\mathbb{P}(\cdot)$, respectively. These are always taken over the 'true' distribution of the data. The symbol $\xrightarrow{p}$ is used to indicate convergence in probability.

We occasionally use the O_p notation, defined as follows. Suppose we have a sequence of random variables, $\{a_n\}$, depending on the sample size n, and a sequence of positive constants, $\{b_n\}$. Then we say that a_n is $O_p(b_n)$ if, for all $\epsilon > 0$, there exist two

constants k_1 and k_2, where k_1 is positive and k_2 depends on k_1, such that

$$\mathbb{P}\left(\left|\frac{a_n}{b_n}\right| > k_1\right) < \epsilon \quad \text{for all} \quad n > k_2.$$

3.1.2 M-estimation theory

The theory of M-estimation will be used later in this chapter to calculate the asymptotic variance of the stratified treatment effect estimator. The relevant ideas are summarized here. For a more comprehensive discussion of M-estimation see, for example, Stefanski and Boos [100] or van der Vaart [108].

Suppose we wish to estimate an unknown — generally vector-valued — population parameter, $\boldsymbol{\theta}_o^{L\times 1}$, for some integer $L \geq 1$. In order to do this a sample of n subjects is drawn from the population and data $\mathbf{W}$ are observed. For example, in a regression model $\mathbf{W}$ would consist of an outcome and a set of covariates and $\boldsymbol{\theta}_o$ would be the vector of regression parameters. In our problem the population stratified treatment effect, β_o^s, will be a component of $\boldsymbol{\theta}_o$.

An M-estimator of $\boldsymbol{\theta}_o$, denoted by $\hat{\boldsymbol{\theta}}$, is obtained as the solution of the estimating equations

$$\sum_{i=1}^{n} \boldsymbol{\psi}\left(\mathbf{W}_i\,;\boldsymbol{\theta}\right) = 0,$$

where $\boldsymbol{\psi}^{L\times 1}$ is a vector of functions of the data and the unknown parameter, and

$$\mathbb{E}_{\mathbf{W}}\left[\,\boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta}_o)\,\right] = 0,$$

where the expectation is taken over the true distribution of the data. Assuming that the estimator $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_o$ and asymptotically follows a multivariate normal (MVN) distribution, we have

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}).$$

Standard M-estimation theory states that this asymptotic covariance matrix, $\boldsymbol{\Sigma}$, can be defined as follows.

$$\boldsymbol{\Sigma}^{L\times L} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-\mathbf{T}},$$

with

$$\mathbf{A}^{L\times L} = \mathbb{E}\left[-\frac{\partial}{\partial\boldsymbol{\theta}^T}\{\boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta})\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\right], \tag{3.1}$$

and

$$\mathbf{B}^{L\times L} = \mathbb{E}\left[\,\boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta}_o)\,\boldsymbol{\psi}^T(\mathbf{W};\boldsymbol{\theta}_o)\,\right].$$

If the component equations of the vector $\boldsymbol{\psi}$ are not smooth in $\boldsymbol{\theta}$ and therefore not differentiable with respect to $\boldsymbol{\theta}$, the matrix $\mathbf{A}$ is undefined. In this situation, the order of differentiation and expectation may be interchanged as follows [100]

$$\mathbf{A}^{L\times L} = \frac{\partial}{\partial\boldsymbol{\theta}^T}\left\{\mathbb{E}\left[-\boldsymbol{\psi}(\mathbf{W}\,;\boldsymbol{\theta})\right]\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}. \tag{3.2}$$

The regularity conditions required for this interchange of differentiation and integration are given, for example, by Huber [43]. We do not state these conditions in full since, for our problem, the proof of normality includes a theorem that ensures the validity of (3.2) (see Appendices A.3 and B.3). When the component equations of the vector $\boldsymbol{\psi}$ are smooth in $\boldsymbol{\theta}$ then the two definitions of $\mathbf{A}$ are equivalent. It is, however, often easier to differentiate first, rather than take expectations and so when both definitions are valid we use (3.1).

Assuming that the appropriate definition of $\mathbf{A}$ is used, the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\,\mathbf{A^{-1}BA^{-T}}.$$

Suppose interest lies only in the first component of the parameter $\boldsymbol{\theta}$ and that the other components are nuisance parameters. The variance of that first component is given by

$$\mathbb{V}\left[\,(\hat{\boldsymbol{\theta}})_1\,\right] = \frac{1}{n}(\mathbf{A^{-1}BA^{-T}})_{11}.$$

3.1.3 *The stratified treatment effect estimator*

We introduced the stratified treatment effect estimator in Chapter 2. We now write this using the new notation (Section 3.1.1). When, as is usual, the propensity score is estimated from the data, $\hat{S}_{si} = 1_{[\hat{q}_{s-1}\le\hat{p}(\mathbf{X})<\hat{q}_s]}$ is an indicator for the s^{th} sample stratum, where $\hat{\mathbf{q}}$ represents the estimated strata boundaries, and $\hat{p}\,(\mathbf{X})$ is the estimated propensity score, and $\hat{\mathbf{d}}$ represents the sample fractions of subjects who are treated

and in each sample stratum, then the stratified treatment effect estimator is

$$\hat{\beta}^s = \frac{1}{n}\sum_{s=1}^{K} r_s \sum_{i=1}^{n} \left\{ \frac{Y_i Z_i \hat{S}_{si}}{\hat{d}_s} - \frac{Y_i (1 - Z_i) \hat{S}_{si}}{r_s - \hat{d}_s} \right\}. \tag{3.3}$$

To calculate $\hat{\beta}^s$ the steps are as follows: (i) estimate the propensity score parameters, $\boldsymbol{\alpha}$; (ii) estimate the strata boundaries, $\mathbf{q}$; (iii) estimate the fraction of subjects who are treated and in each stratum, $\mathbf{d}$; and (iv) using these estimated parameters, calculate the stratified treatment effect estimator (3.3). Note that this procedure assumes that the propensity score is a continuous variable. If the propensity score were discrete, taking finitely many values, then we would define strata by the discrete propensity score, rather than by the quantiles of its distribution.

It is helpful to view each of these estimation steps as the solution to a set of estimating equations. First, the estimated propensity score parameters, $\hat{\boldsymbol{\alpha}}$, can be written as the vector-valued solution to the logistic regression estimating equations. Then given $\hat{\boldsymbol{\alpha}}$, the estimated strata boundaries can be viewed as the vector-valued solution to a further set of estimating equations. In the same way, the fractions of treated subjects in each stratum can be viewed as the vector-valued solution to a set of estimating equations. Finally, we can write $\hat{\beta}^s$ as the solution to an estimating equation involving the sample data and all the other estimated parameters. By combining all the estimating equations, this sequential process of estimation can be viewed as the joint solution of a single set of estimating equations. Thus $\hat{\beta}^s$ is a component of a vector-valued solution to a set of estimating equations — the situation described in Section 3.1.2. After establishing consistency and asymptotic normality, the M-estimation theory can then be applied directly to calculate the variance of the stratified treatment effect estimator. Once we have established consistency and normality, therefore, we proceed to calculate the variance of the stratified treatment effect estimator assuming that the propensity score is: (i) a known function of the observed covariates; and (ii) estimated using a correctly specified logistic regression model. The two resulting variances will be denoted by $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$ where the 'k' and 'e' refer to the propensity score being known and estimated respectively. We shall see that whilst $\mathbb{V}_k[\hat{\beta}^s]$ is typically used in practice we should use $\mathbb{V}_e[\hat{\beta}^s]$, and that the difference between the two variances is often non-trivial.

3.2 Theoretical properties when the propensity score is known

We first consider the simpler hypothetical situation where the propensity score parameters, $\boldsymbol{\alpha}$, are known and so the propensity score is a known function of the observed covariates. In this case, we replace the estimated propensity score, $\hat{p}(\mathbf{X})$, by the known propensity score, $p_o(\mathbf{X})$, in the definition of the stratified treatment effect estimator (3.3).

We derive three sets of estimating equations that are needed to obtain the three sets of estimated parameters, $\hat{\mathbf{q}}, \hat{\mathbf{d}}$ and $\hat{\beta}^s$, where the first two are vectors and the last is a scalar. We will use the estimating equations to demonstrate the consistency and asymptotic normality of each of these estimators. We combine the three sets of estimating equations into a single set of estimating equations where $\hat{\beta}^s$ is the component of interest of the vector-valued solution. We then apply the M-estimation theory (Section 3.1.2) to obtain the asymptotic variance of the stratified treatment effect estimator. Next, we therefore determine the three sets of estimating equations for $\hat{\mathbf{q}}$, $\hat{\mathbf{d}}$ and $\hat{\beta}^s$.

Estimating the strata boundaries

We first wish to find a set of estimating equations that, when solved, estimate the strata boundaries, which are defined as quantiles of the population distribution of the propensity score. We write the propensity score as $p_o(\mathbf{X})$ since we are assuming that the propensity score parameters are known. The estimation of the the strata boundaries is equivalent to solving the estimating equations $\sum_{i=1}^{n} \boldsymbol{\psi}_3 (\mathbf{X}_i ; \mathbf{q}) = 0$ for $\mathbf{q}$, where, remembering that the stratum indicator $S_{si} = 1_{[q_{s-1} \leq p_o(\mathbf{X}_i) < q_s]}$ is a function of $\mathbf{q}$,

$$\boldsymbol{\psi}_3^{(K-1)\times 1} (\mathbf{X}_i ; \mathbf{q}) = \begin{pmatrix} S_{1i} - r_1 \\ \vdots \\ S_{(K-1)i} - r_{(K-1)} \end{pmatrix}.$$

We define the strata boundaries q_0 and q_K as 0 and 1, respectively. The fractions of the sample in each stratum, $r_1, r_2, ..., r_{(K-1)}$, are fixed by the analyst. Then the equation

$$\sum_{i=1}^{n} (S_{1i} - r_1) = \sum_{i=1}^{n} \left(1_{[0 \leq p_o(\mathbf{X}_i) < q_1]} - r_1 \right) = 0$$

has only one unknown: q_1. Solving this equation determines $\hat{q}_1$, the estimate of the first strata boundary, as the value of the propensity score such that exactly $n\,r_1$ subjects in the sample have a propensity score of less than $\hat{q}_1$ [1]. Given $\hat{q}_1$, the estimate of the second strata boundary is determined as the solution, $\hat{q}_2$, of

$$\sum_{i=1}^{n}(S_{2i} - r_2) = \sum_{i=1}^{n}\left(1_{[\,\hat{q}_1 \le p_o(\mathbf{X}_i) < q_2\,]} - r_2\right) = 0.$$

Proceeding in this way, the estimating equations $\sum_{i=1}^{n}\psi_3\,(\mathbf{X}_i\,;\mathbf{q}) = 0$ estimate the quantiles of the propensity score distribution in the population by the quantiles of the propensity score distribution in the sample.

Estimating the fraction treated in each stratum

Given the estimated strata boundaries, we now wish to find a set of estimating equations that, when solved, estimate the fraction of subjects who are treated and in each stratum. Remembering that Z_i denotes the treatment status of subject i and, given the estimated strata boundaries $\hat{\mathbf{q}}$, the sample stratum indicator is defined as $\hat{S}_{si} = 1_{[\hat{q}_{s-1} \le p_o(\mathbf{X}) < \hat{q}_s]}$, the required fractions can be estimated by solving the estimating equations $\sum_{i=1}^{n}\psi_2\,(Z_i, \mathbf{X}_i\,;\mathbf{d}, \hat{\mathbf{q}}) = 0$ for $\mathbf{d}$, with

$$\psi_2^{K\times 1}\,(Z_i, \mathbf{X}_i\,;\mathbf{d}, \hat{\mathbf{q}}) = \begin{pmatrix} Z_i\;\hat{S}_{1i} - d_1 \\ \vdots \\ Z_i\;\hat{S}_{Ki} - d_K \end{pmatrix}.$$

To see this, note that given the estimated strata boundaries, $\hat{\mathbf{q}}$, the only unknown in the equation

$$\sum_{i=1}^{n}\left(Z_i\;\hat{S}_{1i} - d_1\right) = \sum_{i=1}^{n}\left(Z_i\;1_{[\,0 \le p_o(\mathbf{X}_i) < \hat{q}_1\,]} - d_1\right) = 0$$

is d_1. Solving the equation determines $\hat{d}_1$ as the fraction of subjects who are treated and who are in the first sample stratum. Given this estimate, $\hat{d}_2$ is then determined as the fraction of the sample who are treated and in the second sample stratum. Continuing in this way, the population probability of being both treated and in the

[1] In practice, some nr_s may not be integers, in which case we choose the strata boundaries to approximately satisfy this condition. In ascertaining the theoretical properties of the estimators we assume that all the estimating equations can be exactly solved. A small approximation, in practice, however, will not materially affect the variance, consistency or normality of the stratified treatment effect estimator.

s^{th} stratum, d_{so}, is estimated by the sample fraction of subjects who are both treated and in the s^{th} sample stratum, $\hat{d}_s$, for strata $s = 1, \ldots . K$.

Estimating the stratified treatment effect

Having estimated both the strata boundaries and the probabilities of being treated and in each stratum, the stratified treatment effect estimator (3.3) can be calculated by solving the equation

$$\sum_{i=1}^{n} \psi_1 \left(Y_i, Z_i, \mathbf{X}_i \, ; \beta^s, \hat{\mathbf{d}}, \hat{\mathbf{q}}\right) = 0, \tag{3.4}$$

for β^s, where, remembering that Y_i and Z_i denote the outcome and treatment status of subject i, and $\hat{S}_{si} = 1_{[\hat{q}_{s-1} \leq p_o(\mathbf{X}) < \hat{q}_s]}$ is the sample estimate of the s^{th} stratum indicator,

$$\psi_1^{1\times 1} \left(Y_i, Z_i, \mathbf{X}_i \, ; \beta^s, \hat{\mathbf{d}}, \hat{\mathbf{q}}\right) = \sum_{s=1}^{K} r_s \left\{ \frac{Y_i \, Z_i \, \hat{S}_{si}}{\hat{d}_s} - \frac{Y_i \, (1 - Z_i) \, \hat{S}_{si}}{r_s - \hat{d}_s} \right\} - \beta^s.$$

Solving the estimating equation (3.4) determines $\hat{\beta}^s$ as

$$\hat{\beta}^s = \frac{1}{n} \sum_{s=1}^{K} r_s \sum_{i=1}^{n} \left\{ \frac{Y_i \, Z_i \, \hat{S}_{si}}{\hat{d}_s} - \frac{Y_i \, (1 - Z_i) \, \hat{S}_{si}}{r_s - \hat{d}_s} \right\}.$$

A single set of estimating equations

We have described how the stratified treatment effect estimator can be obtained by the sequential solution of the three sets of estimating equations discussed above. We can alternatively consider the simultaneous estimation of a vector of unknown parameters, $\boldsymbol{\theta}_o^{2K\times 1}$, defined by $\boldsymbol{\theta}_o^T = (\beta_o^s, \mathbf{d}_o^T, \mathbf{q}_o^T)$. Essentially, if we 'stack' the three sets of estimating equations on top of one another and solve them simultaneously, it is equivalent to the sequential solution described above. In particular, if we define a vector $\boldsymbol{\psi}^{2K\times 1}$ by $\boldsymbol{\psi}^T = (\psi_1, \boldsymbol{\psi}_2^T, \boldsymbol{\psi}_3^T)$, then solving the set of estimating equations

$$\sum_{i=1}^{n} \boldsymbol{\psi} \left(Y_i, Z_i, \mathbf{X}_i \, ; \boldsymbol{\theta}\right) = 0, \tag{3.5}$$

for $\boldsymbol{\theta}$ gives an estimator, $\hat{\boldsymbol{\theta}}$, defined by $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T)$. To see that the two approaches are equivalent, note that working from the bottom of the new vector of estimating equations upwards, each component uniquely determines a component of $\hat{\boldsymbol{\theta}}$, given the previous estimates.

We have now done what we set out to do at the beginning of this section. We use this representation of the estimation process to demonstrate the consistency and asymptotic normality of the stratified treatment effect estimator. We will then be in a position to apply the M-estimation theory to calculate the variance of this estimator.

3.2.1 Consistency, asymptotic normality and variance

We have seen that, assuming that the propensity score is a known function of the observed covariates, the stratified treatment effect estimator, $\hat{\beta}^s$, can be calculated as the first component of a vector of estimated parameters, $\hat{\boldsymbol{\theta}}$, obtained by solving the joint estimating equation (3.5). We now investigate the theoretical properties of $\hat{\beta}^s$.

Intuitively, we would expect the estimated strata boundaries, $\hat{\mathbf{q}}$, to be consistent estimates of the population strata boundaries, $\mathbf{q}_o$, and the estimated fractions of treated subjects in each stratum, $\hat{\mathbf{d}}$, to be consistent estimates of the equivalent population probabilities, $\mathbf{d}_o$. We would also expect the stratified treatment effect estimator, $\hat{\beta}^s$, to be a consistent estimate of its 'true' value, β_o^s,

$$\beta_o^s = \sum_{s=1}^{K} r_s \{ \mathbb{E}[\, Y \mid Z = 1,\, S_{so} = 1\,] - \mathbb{E}[\, Y \mid Z = 0,\, S_{so} = 1\,] \}.$$

The following lemma establishes conditions under which the three sets of estimators, $\hat{\mathbf{q}}$, $\hat{\mathbf{d}}$ and $\hat{\beta}^s$, are consistent estimators of the population parameters mentioned above. These conditions are not minimal, but they are sufficient for consistency and turn out to be necessary for asymptotic normality.

Lemma 3.1 *Suppose that the propensity score is a known function of the observed covariates.*

(i) Suppose that the cumulative distribution function of the propensity score is strictly monotone and continuous near the population strata boundaries, $\mathbf{q}_o$. *Then as* $n \to \infty$, $\hat{\mathbf{q}} \xrightarrow{p} \mathbf{q}_o$.

(ii) Suppose that condition (i) is satisfied, and so $\hat{\mathbf{q}}$ *is consistent. Suppose further that the probability density function of the propensity score,* $f_p(\cdot)$, *is bounded near the population strata boundaries and that the population probability of being*

treated and in the s^{th} stratum, d_{so}, is not equal to either 0 or r_s, for $s = 1..., K$. Then as $n \to \infty$, $\hat{\mathbf{d}} \xrightarrow{p} \mathbf{d}_o$.

(iii) Suppose that conditions (i) and (ii) are satisfied, and so both $\hat{\mathbf{q}}$ and $\hat{\mathbf{d}}$ are consistent. Suppose further that the following functions are bounded for all $p \in (0,1)$ and $t = 0, 1$:

$$f_p(p), \qquad \mathbb{E}[Y \mid Z = t, p_o(\mathbf{X}) = p], \qquad \mathbb{E}[Y^2 \mid Z = t, p_o(\mathbf{X}) = p].$$

Then as $n \to \infty$, $\hat{\beta}^s \xrightarrow{p} \beta_o^s$.

If all the above conditions are satisfied then as $n \to \infty$, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_o$, where $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T)$ and $\boldsymbol{\theta}_o^T = (\beta_o^s, \mathbf{d}_o^T, \mathbf{q}_o^T)$.

◇

Proof of this lemma can be found in Appendix A.2. The following theorem establishes conditions under which $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed and calculates the asymptotic variance of the stratified treatment effect estimator, $\hat{\beta}^s$.

Theorem 3.1 *Suppose that the propensity score is a known function of the observed covariates and that:*

(i) the conditions of Lemma 3.1 are satisfied;

(ii) the probability density function of the propensity score, $f_p(\cdot)$, is non-zero at each population strata boundary, q_{so}, for $s = 1, ..., K-1$;

(iii) the functions $\mathbb{E}[Y \mid Z = t, p_o(\mathbf{X}) = p]$ are continuous in p near the population strata boundaries, for $t = 0, 1$.

Then $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed, where $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T)$. Furthermore, the asymptotic variance of the stratified treatment effect estimator, $\hat{\beta}^s$, is

$$\mathbb{V}_k[\hat{\beta}^s] = V_1 + V_2,$$

with

$$V_1 = \frac{1}{n} \sum_{s=1}^{K} r_s^2 \left\{ \frac{\mathbb{V}[Y \mid Z = 1, S_{so} = 1]}{d_{so}} + \frac{\mathbb{V}[Y \mid Z = 0, S_{so} = 1]}{r_s - d_{so}} \right\},$$

$$V_2 = \frac{1}{n} \left. \frac{\partial \beta^*}{\partial \mathbf{q}^T} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}_o} (n Cov[\hat{\mathbf{q}}]) \left. \frac{\partial \beta^*}{\partial \mathbf{q}} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}_o},$$

where $(n\ Cov[\hat{\mathbf{q}}])$ *is a* $(K-1)\times(K-1)$ *matrix representing the asymptotic covariance matrix of the estimated strata boundaries, defined for* $j,k=1,....,K-1,\ j\geq k$*. as*

$$(n\ Cov[\,\hat{\mathbf{q}}\,])_{jk}=\frac{\mathbb{P}(p_o(\mathbf{X})>q_{jo})\ \mathbb{P}(p_o(\mathbf{X})<q_{ko})}{f_p(q_{jo})\ f_p(q_{ko})},$$

and $\left.\frac{\partial\beta^*}{\partial\mathbf{q}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$ *is a* $1\times(K-1)$ *vector with*

$$\beta^*=\sum_{s=1}^{K}r_s\{\mathbb{E}\,[\,Y\mid Z=1,S_s=1\,]-\mathbb{E}\,[\,Y\mid Z=0,S_s=1\,]\},$$

which is equal to the 'true' value of $\hat{\beta}^s$*,* β_o^s*, but seen as a function of the strata boundaries,* $\mathbf{q}$*, rather than evaluated at the population strata boundaries.*

◇

Proof of this theorem can be found in Appendices A.3 and A.4.

We now discuss the practical implications of the conditions in Lemma 3.1 and Theorem 3.1. Condition *(i)* of Lemma 3.1 and condition *(ii)* of Theorem 3.1 both ensure that the strata boundaries are well defined and can therefore be estimated. Later on, we will see an example which almost violates one of these conditions — the probability density function is extremely close to zero at one of the population strata boundaries (Section 5.2.4). In this example the empirical sampling distribution of that estimated strata boundary is non-normal, and the empirical variance of the stratified treatment effect estimator is not equal to $\mathbb{V}_k\,[\hat{\beta}^s]$. It is worth noting that the variance $\mathbb{V}_k\,[\hat{\beta}^s]$ is an asymptotic result, due to the implicit reliance of the variance calculation on the central limit theorem. Therefore, when the conditions in Lemma 3.1 and Theorem 3.1 are satisfied, the variance formula $\mathbb{V}_k\,[\hat{\beta}^s]$ is valid for 'large enough' samples but the sample size required to achieve this asymptotic variance will increase the closer we come to violating the conditions.

Condition *(ii)* of Lemma 3.1 states that that the population probabilities of being treated and in each stratum, d_{so}, must not be 0 or r_s, for $s=1,...,K$, which ensures that the estimand β_o^s is well defined. It is easy to imagine a situation where this condition is violated. As an example, let us consider the use of statins to lower serum cholesterol levels. These are given to patients with high cholesterol who also have an increased risk of a cardiovascular event, such as a prior myocardial infarction, angina

or characteristics including high blood pressure and large waist circumference. It may, however, be extremely hard to find subjects with these characteristics — and hence a high propensity score — who are not treated with statins. Then a stratified propensity score analysis might create strata in which the higher strata contain only patients taking statins and the lower strata contain only patients not taking statins. In this case, the stratified treatment effect estimator will be undefined. This problem will be apparent when the data is analysed using stratification on the propensity score, and is typically dealt with by widening the strata boundaries, although this often means that the treated and untreated groups within the strata are then not comparable.

The remaining conditions demand that the probability density function of the propensity score, the conditional expectation of the outcome given the propensity score, and the conditional squared expectation of the outcome given the propensity score are all bounded. It is hard to imagine a practical situation where this is not the case. However, it is easy to imagine situations where the probability density function is very large, for example when all subjects have a very similar propensity score. Then, although the condition is not absolutely violated, we might require a large sample size for the variance formula $\mathbb{V}_k[\,\beta^s\,]$ to be valid. In fact, we will see a simulated example where this is the case later on (Section 5.2.3).

3.3 Theoretical properties when the propensity score is estimated

We have considered the theoretical properties of the stratified treatment effect estimator when the propensity score is a known function of the observed covariates. In applications, the propensity score is invariably unknown and therefore must be estimated from the data. In epidemiological applications, this estimation is typically performed using logistic regression of the treatment indicator, Z, on the observed covariates $\mathbf{X}$. This assumes the following relationship,

$$\ln\left\{\frac{p_o(\mathbf{X})}{1-p_o(\mathbf{X})}\right\} = \boldsymbol{\alpha}_o^T\mathbf{X}.$$

We assume that the logistic regression model above is correctly specified. Fitting this model gives us our estimated propensity score parameters, $\hat{\boldsymbol{\alpha}}$, which are used to estimate the propensity score for each sampled subject. Then we calculate the stratified treatment effect estimator as we did when the propensity score was known.

When the propensity score is estimated, we can obtain the stratified treatment effect estimator through the sequential sequence of four sets of estimating equations. The first is the standard set of estimating equations for a logistic regression model. The other three are ψ_3, ψ_2 and ψ_1 — the estimating equations that we solve to obtain the stratified treatment effect estimator when the propensity score is known. We combine these four sets of estimating equations in order to obtain a single set of estimating equations where $\hat{\beta}^s$ is the component of interest of the vector-valued solution. We can then apply the M-estimation theory (Section 3.1.2) to obtain the asymptotic variance of the stratified treatment effect estimator. Next, we therefore derive the fourth estimating equation that is needed to estimate the propensity score.

Estimating the propensity score parameters

For the estimation of the propensity score, we assume that there are m observed covariates including an 'intercept' vector of 1's, $\mathbf{X} = (X_1, X_2, ..., X_m)$. Fitting a logistic regression model with these covariates is equivalent to solving the estimating equations $\sum_{i=1}^{n} \boldsymbol{\psi}_4 \left(Z_i, \mathbf{X}_i \, ; \boldsymbol{\alpha}\right) = 0$ where

$$\boldsymbol{\psi}_4^{m\times 1} \left(Z_i, \mathbf{X}_i \, ; \boldsymbol{\alpha}\right) = \begin{pmatrix} \left(Z_i - \frac{\exp\left(\boldsymbol{\alpha}^T \mathbf{X}_i\right)}{1+\exp\left(\boldsymbol{\alpha}^T \mathbf{X}_i\right)}\right) X_{1i} \\ \vdots \\ \left(Z_i - \frac{\exp\left(\boldsymbol{\alpha}^T \mathbf{X}_i\right)}{1+\exp\left(\boldsymbol{\alpha}^T \mathbf{X}_i\right)}\right) X_{mi} \end{pmatrix},$$

where X_{1i} denotes the observed value of covariate X_1 for subject i. These estimating equations, $\sum_{i=1}^{n} \boldsymbol{\psi}_4 \left(Z_i, \mathbf{X}_i \, ; \boldsymbol{\alpha}\right) = 0$, are the usual score functions for a maximum likelihood logistic regression model and so solving them gives the maximum likelihood estimator of $\boldsymbol{\alpha}_o$. Once the propensity score parameters have been estimated, the propensity score can be treated as if it were a known function of the observed covariates and the stratified treatment effect estimator can be obtained as before.

A single set of estimating equations

The stratified treatment effect estimator can be obtained by the sequential solution of four sets of estimating equations. We can alternatively consider the simultaneous estimation of a vector of unknown parameters, $\boldsymbol{\theta}_o^{(2K+m)\times 1}$, defined by $\boldsymbol{\theta}_o^T = (\beta_o^s, \mathbf{d}_o^T, \mathbf{q}_o^T, \boldsymbol{\alpha}_o^T)$. As before, if we 'stack' the four sets of estimating equations on top of one another and solve them simultaneously, it is equivalent to the

sequential solution described above. In particular, if we define a vector $\boldsymbol{\psi}^{(2K+m)\times 1}$ by $\boldsymbol{\psi}^T = (\psi_1, \boldsymbol{\psi}_2^T, \boldsymbol{\psi}_3^T, \boldsymbol{\psi}_4^T)$, then solving the set of estimating equations

$$\sum_{i=1}^{n} \boldsymbol{\psi}\left(Y_i, Z_i, \mathbf{X}_i ; \boldsymbol{\theta}\right) = 0, \tag{3.6}$$

for $\boldsymbol{\theta}$ gives an estimator, $\hat{\boldsymbol{\theta}}$, defined by $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T, \hat{\boldsymbol{\alpha}}^T)$. This is equivalent to the sequential solution of the four separate sets of estimating equations and results in the same estimators.

We have now done what we set out to do at the beginning of this section. We use this representation of the estimation process to demonstrate the consistency and asymptotic normality of the stratified treatment effect estimator when the propensity score is estimated. We will then be in a position to apply the M-estimation theory to calculate the variance of the estimator.

3.3.1 Consistency, asymptotic normality and variance

We have seen that, assuming that the propensity score is estimated using a correctly specified logistic regression model, the stratified treatment effect estimator, $\hat{\beta}^s$, can be calculated as the first component in a vector of estimated parameters, $\hat{\boldsymbol{\theta}}$, obtained by solving the joint estimating equation (3.6). We now investigate the theoretical properties of $\hat{\beta}^s$.

The following lemma establishes conditions under which the four sets of estimators, $\hat{\boldsymbol{\alpha}}$, $\hat{\mathbf{q}}, \hat{\mathbf{d}}$ and $\hat{\beta}^s$ are consistent. These conditions are not minimal, but as before they are sufficient for consistency and turn out to be necessary for asymptotic normality.

Lemma 3.2 *Suppose that the propensity score is estimated using a correctly specified logistic regression model.*

(i) Suppose that the propensity score parameters, $\hat{\boldsymbol{\alpha}}$, are consistent and that the cumulative distribution function of the propensity score is strictly monotone and continuous near the population strata boundaries, $\mathbf{q}_o$. Suppose also that the probability density function of the propensity score is continuous everywhere, and that the derivatives $\frac{\partial f_p(\cdot)}{\partial \alpha_k}$ exist and are bounded, for $k = 1, \ldots, m$.
Then as $n \to \infty$, $\hat{\mathbf{q}} \xrightarrow{p} \mathbf{q}_o$.

(ii) Suppose that $\hat{\boldsymbol{\alpha}}$ *is consistent and that condition (i) is satisfied, so* $\hat{\mathbf{q}}$ *is consistent. Suppose further that the probability density function of the propensity score is bounded near the population strata boundaries, and that the population probability of being treated and stratum* s, d_{so}, *is not equal to either 0 or* r_s, *for* $s = 1, .., K$.
Then as $n \to \infty$, $\hat{\mathbf{d}} \xrightarrow{p} \mathbf{d}_o$.

(iii) Suppose that $\hat{\boldsymbol{\alpha}}$ *is consistent and that conditions (i) and (ii) are satisfied, so both* $\hat{\mathbf{q}}$ *and* $\hat{\mathbf{d}}$ *are consistent. Suppose further that the following functions exist and are continuous in* p *and bounded everywhere, for* $t = 0, 1$ *and* $k = 1, ..., m$:

$$\mathbb{E}[Y \mid Z = t,\, p_o(\mathbf{X}) = p], \qquad \mathbb{E}[Y^2 \mid Z = t,\, p_o(\mathbf{X}) = p],$$
$$\frac{\partial}{\partial \alpha_k}\left\{\mathbb{E}[Y \mid Z = t,\, p(\mathbf{X}) = p]\right\}, \qquad \frac{\partial}{\partial \alpha_k}\left\{\mathbb{E}[Y^2 \mid Z = t,\, p(\mathbf{X}) = p]\right\}.$$

Then as $n \to \infty$, $\hat{\beta}^s \xrightarrow{p} \beta_o^s$.

If all the above conditions are satisfied then as $n \to \infty$, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_o$, *where* $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T, \hat{\boldsymbol{\alpha}}^T)$ *and* $\boldsymbol{\theta}_o^T = (\beta_o^s, \mathbf{d}_o^T, \mathbf{q}_o^T, \boldsymbol{\alpha}_o^T)$.

◇

Proof of this Lemma can be found in Appendix B.2. The following theorem now establishes conditions under which $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed and calculates the asymptotic variance of the stratified treatment effect estimator, $\hat{\beta}^s$.

Theorem 3.2 *Suppose that the propensity score is estimated using a correctly specified logistic regression model and that:*

(i) the conditions of Lemma 3.2 are satisfied;

(ii) the probability density function of the propensity score is non-zero at each population strata boundary, q_{so}, *for* $s = 1, ..., K-1$, *but zero at 0 and 1.*

Then $\hat{\boldsymbol{\theta}}$ *is asymptotically normally distributed, where* $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T, \boldsymbol{\alpha}^T)$. *Furthermore, the asymptotic variance of the stratified treatment effect estimator,* $\hat{\beta}^s$, *is*

$$\mathbb{V}_e[\hat{\beta}^s] = V_1 + V_2 + V_3 + V_4,$$

where V_1 *and* V_2 *are defined as in Theorem 3.1, and*

$$V_3 = -\frac{1}{n}\,\mathbf{C}\,(n\;Cov[\,\hat{\boldsymbol{\alpha}}\,])\,\mathbf{C}^T$$
$$V_4 = \quad\frac{1}{n}\quad\mathbf{e}\,(n\;Cov[\,\hat{\boldsymbol{\alpha}}\,])\quad\mathbf{e}^T$$

where $\mathbf{C} = (C_1, ..., C_m)$ *is defined, for* $k = 1, ..., m$, *as*

$$C_k = \sum_{s=1}^{K} r_s\;Cov[\,Y,\,X_k\,(1 - p_o(\mathbf{X}))\,|\,Z = 1,\,S_{so} = 1\,]$$
$$+ \quad\sum_{s=1}^{K} r_s\;Cov[\,Y,\,X_k\,p_o(\mathbf{X})\,|\,Z = 0,\,S_{so} = 1\,].$$

The covariance matrix $(n\;Cov[\,\hat{\boldsymbol{\alpha}}\,])$ *is a* $m \times m$ *matrix defined in terms of its inverse, for* $j, k = 1, ..., m$, *as*

$$(n\;\;Cov[\,\hat{\boldsymbol{\alpha}}\,])^{-1}_{jk} = \mathbb{E}\,[\,p_o(\mathbf{X})\,(1 - p_o(\mathbf{X}))\;\;X_j\;\;X_k\,].$$

Finally, $\mathbf{e} = (e_1, ..., e_m)$, *for* $k = 1, ..., m$, *is defined as* $e_k = e_{\boldsymbol{\alpha}k} + e_{\mathbf{q}k}$, *with*

$$e_{\boldsymbol{\alpha}k} = \sum_{s=1}^{K} r_s\,\left\{\frac{(I_{Y_1k} - \mathbb{E}\,[\,Y\,|\,Z = 1, S_{so} = 1\,]\,I_{f_1k})}{d_{so}} - \frac{(I_{Y_0k} - \mathbb{E}\,[\,Y\,|\,Z = 0, S_{so} = 1\,]\,I_{f_0k})}{r_s - d_{so}}\right\},$$

where

$$I_{f_1k} = \int_{q_{(s-1)o}}^{q_{so}} r\frac{\partial}{\partial\alpha_k}\,\{f_p(r;\boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\;\;dr$$

$$I_{f_0k} = \int_{q_{(s-1)o}}^{q_{so}} (1-r)\frac{\partial}{\partial\alpha_k}\,\{f_p(r;\boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\;\;dr$$

$$I_{Y_1k} = \int_{q_{(s-1)o}}^{q_{so}} r\frac{\partial}{\partial\alpha_k}\,\{\mathbb{E}\,[\,Y_1\,|\,p\,(\mathbf{X};\boldsymbol{\alpha}) = r\,]\,f_p(r;\boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\;\;dr$$

$$I_{Y_0k} = \int_{q_{(s-1)o}}^{q_{so}} (1-r)\frac{\partial}{\partial\alpha_k}\,\{\mathbb{E}\,[\,Y_0\,|\,p\,(\mathbf{X};\boldsymbol{\alpha}) = r\,]\,f_p(r;\boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\;\;dr,$$

and

$$e_{\mathbf{q}k} = \sum_{j=1}^{K-1} \left.\frac{\partial\beta^*}{\partial q_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} f_p(q_{jo})^{-1}\;\;\frac{\partial}{\partial\alpha_k}\,\{\mathbb{E}\,[\,1_{[p\,(\mathbf{X};\boldsymbol{\alpha})<q_j]}\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

◇

Proof of the theorem above is given in Appendices B.3 and B.4.

We have already discussed the implications of most of the conditions contained in Lemma 3.2 and Theorem 3.2. The additional ones are concerned with the existence of derivatives, taken with respect to the propensity score parameters. These will exist if the derivatives taken with respect to the propensity score exist. A derivative will not exist at points where, for example, the function has a discontinuity, or has a 'kink', or tends to infinity. Clinical examples where either the probability density function of the propensity score or the conditional expectation of the outcome, given treatment status and propensity score, tends to infinity are hard to imagine. Such problems, moreover, would be apparent when carrying out the analysis on a dataset. Discontinuities and 'kinks' in these functions are more likely to occur.

One example of a discontinuity in the probability density function of the propensity score is the BCG vaccination in Britain. This vaccination is given to schoolchildren aged 14 years old. Therefore, the propensity score for this treatment, the vaccine, is merely a function of age. Furthermore, if schools gave the vaccine to each child exactly on their 14^{th} birthday then the cumulative density function of the propensity score would be

$$F_p(p) = \begin{cases} 0 & \text{if age} < 14 \\ 1 & \text{if age} \geq 14 \end{cases}$$

This gives a non-continuous probability density function. However, since schools do not vaccinate each child so promptly after their 14^{th} birthday it is likely that the true propensity score here is continuous.

Similarly, we can find examples of discontinuity of the conditional expectation of the outcome, given treatment status and propensity score. Suppose we wished to estimate the effect of a particular intervention that aimed to curb binge drinking in schoolchildren. Suppose we targeted the intervention primarily at the older schoolchildren, so the propensity score depended only on age. Then if these particular children lived in an area where it was hard to obtain alcohol illegally, we would expect an increase in binge drinking, the outcome, immediately after a child's 18^{th} birthday, creating a discontinuity. However, in practice this problem would not occur since we would use

our knowledge of the situation to categorize the children into two groups — under 18 and over 18.

Therefore, although it is fairly easy to think of examples in which these conditions would be violated, these examples are unlikely to occur in practice. Conversely, practical examples where these conditions are *almost* violated may be much more common.

3.4 Components of variability of the stratified treatment effect estimator

We have now ascertained various theoretical properties of the stratified treatment effect estimator, $\hat{\beta}^s$. In particular, we have calculated, using M-estimation methods, the variance of $\hat{\beta}^s$ assuming that the propensity score is: (i) a known function of the observed covariates; and (ii) estimated by a correctly specified logistic regression model. The two variance formulæ are denoted by $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$, where the k and e subscripts refer to the propensity score being 'known' and 'estimated' respectively. These variances can be written in terms of four variance components as follows:

$$\mathbb{V}_k[\hat{\beta}^s] = V_1 + V_2,$$
$$\mathbb{V}_e[\hat{\beta}^s] = V_1 + V_2 + V_3 + V_4.$$

We now consider the interpretation of V_1, V_2, V_3 and V_4.

3.4.1 The variance component V_1

If, as before, we let K denote the number of strata, r_s denote the probability of being in the s^{th} stratum, d_{so} be the population probability of being treated and in the s^{th} stratum, q_{so} be the s^{th} population strata boundary, and $S_{so} = 1_{[q_{(s-1)o} \leq p_o(\mathbf{X}) < q_{so}]}$ be an indicator for the s^{th} population stratum, then, from Theorem 3.1, the first variance component is

$$V_1 = \frac{1}{n} \sum_{s=1}^{K} r_s^2 \left\{ \frac{\mathbb{V}[Y \mid Z = 1, S_{so} = 1]}{d_{so}} + \frac{\mathbb{V}[Y \mid Z = 0, S_{so} = 1]}{r_s - d_{so}} \right\}.$$

Since this variance component consists of the sum of variances, inversely weighted by positive probabilities, it is always positive. In the hypothetical situation where both the propensity score and the strata boundaries are known, and sampling is performed within strata, the variance of the stratified treatment effect estimator is exactly equal to V_1. This variance component therefore measures the error due to the variability of the outcome, Y, within the true strata. Since this error occurs whether the propensity score is known or estimated, both $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$ contain this term. In all the simulated examples we consider later (Section 5.2), the component V_1 accounts for most of the variance of the stratified treatment effect estimator.

3.4.2 The variance component V_2

From Theorem 3.1, the second variance component is

$$V_2 = \frac{1}{n}\left.\frac{\partial\beta^*}{\partial\mathbf{q}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} (n\mathrm{Cov}[\,\hat{\mathbf{q}}\,]) \left.\frac{\partial\beta^*}{\partial\mathbf{q}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o},$$

where $(n\ \mathrm{Cov}[\hat{\mathbf{q}}])$ is a $(K-1)\times(K-1)$ matrix representing the asymptotic covariance matrix of the estimated strata boundaries, defined for $j,k = 1,...,K-1,\ j \geq k$, as

$$(n\ \mathrm{Cov}[\,\hat{\mathbf{q}}\,])_{jk} = \frac{\mathbb{P}(p_o(\mathbf{X}) > q_{jo})\ \mathbb{P}(p_o(\mathbf{X}) < q_{ko})}{f_p(q_{jo})\ f_p(q_{ko})}, \tag{3.7}$$

and $\left.\frac{\partial\beta^*}{\partial\mathbf{q}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$ is a $1\times(K-1)$ vector with

$$\beta^* = \sum_{s=1}^{K} r_s\{\mathbb{E}\,[\,Y \mid Z=1, S_s=1\,] - \mathbb{E}\,[\,Y \mid Z=0, S_s=1\,]\}.$$

Since the variance component V_2 is a quadratic form centred around a positive definite matrix, it is always positive. The quantiles of the propensity score distribution — the strata boundaries — must be estimated whether the propensity score is known or estimated and so both $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$ contain this term. Thus the estimation of the strata boundaries will always increase the variance of the stratified treatment effect estimator, although, as the following discussion will show, this increase can be expected to be negligible. We find that the component V_2 is negligible in all the simulated examples considered later (Section 5.2).

We now consider situations in which we might expect V_2 to be large. We begin by considering the derivative term that appears either side of the covariance matrix in V_2. As mentioned previously, the quantity β^* is the 'true' value of $\hat{\beta}^s$, β^s_o, seen as a function of the strata boundaries, $\mathbf{q}$, rather than evaluated at the population strata boundaries. The derivative of β^* with respect to the strata boundaries, therefore, describes the change in the population quantity being estimated caused by a change in the population strata boundaries. This term will be large if a small change in a strata boundary causes a large change in the treatment effect estimand. In practice, we would not expect this to be the case and so the derivative terms in V_2 will typically be small.

The asymptotic covariance matrix of the estimated strata boundaries (3.7) will only be large if the probability density function of the propensity score is extremely low at the population strata boundaries, in line with our intuition that estimates of a quantile in an area with few observations will be very variable. In Section 5.2.4 we will meet a simulated example where this occurs. However, even when the components of the covariance matrix are large, the components of the derivative are likely to be small so we would expect V_2 to contribute little to the overall variance, as is the case in the simulated example mentioned above.

Note that one of the conditions for the validity of both the asymptotic variance formulæ is that the probability density function of the propensity score is non-zero at the population strata boundaries. Therefore, if the probability density function of the propensity score is very low at a population strata boundary a bigger sample size will be needed for normality of the stratified treatment effect estimator and the validity of the variance formulæ — possibly prohibitively bigger. This occurs in the example mentioned previously in Section 5.2.4 where we give some practical insight into reasons for this problem.

3.4.3 The variance component V_3

From Theorem 3.2, the third variance component is

$$V_3 = -\frac{1}{n}\ \mathbf{C}\ (n\ \mathrm{Cov}[\,\hat{\boldsymbol{\alpha}}\,])\ \ \mathbf{C}^T.$$

Recalling that m is the number of covariates used to estimate the propensity score (including a constant vector) we define $\mathbf{C} = (C_1, C_2, ..., C_m)$, where for $k = 1, ..., m$,

$$C_k = \sum_{s=1}^{K} r_s \, \mathrm{Cov}[\, Y,\ X_k \,(1 - p_o(\mathbf{X}))\,|\, Z = 1,\ S_{so} = 1\,] + \sum_{s=1}^{K} r_s \, \mathrm{Cov}[\, Y,\ X_k \, p_o(\mathbf{X})\,|\, Z = 0,\ S_{so} = 1\,].$$

The covariance matrix is defined, for $j, k = 1, ..., m$, as

$$(n \, \mathrm{Cov}[\, \hat{\boldsymbol{\alpha}}\,])^{-1}_{jk} = \mathbb{E}\,[\, p_o(\mathbf{X})\,(1 - p_o(\mathbf{X}))\, X_j \, X_k\,]. \tag{3.8}$$

The variance component V_3 is also a quadratic form around a positive definite matrix. Since this quadratic form is preceded by a minus sign, V_3 will always be negative. As we will see, this term measures the extent to which the estimation of the propensity score reduces the variance of the stratified treatment effect estimator by increasing the balance of the distributions of covariates that are associated with outcome within the strata. Conversely, the estimation of the propensity score also introduces random error which increases the variance of the stratified treatment effect estimator, measured by the positive variance component V_4. In each simulated example considered later (Section 5.2) the overall effect of estimating the propensity score, measured by $V_3 + V_4$, is to decrease the variance.

The matrix (3.8) at the heart of the component V_3 is the asymptotic covariance matrix of the estimated propensity score parameters, $\hat{\boldsymbol{\alpha}}$. This is the standard covariance matrix for logistic regression model parameters. We can see that this matrix will be large when many of the propensity score values are very close to either zero or one.

The vector $\mathbf{C}$ is a measure of the covariance of the outcome, Y, and the covariates, $\mathbf{X}$, weighted by a function of the propensity score. This term will increase as the covariance between Y and $\mathbf{X}$ increases. In order to better understand this term, we consider a much simpler situation. Suppose that each population stratum only contains a single value of the propensity score, and that the individual level treatment effect is the same for each subject. Then

$$C_k = \sum_{s=1}^{K} r_s \, \mathrm{Cov}[\, Y_1,\ X_k \,|\, p_o(\mathbf{X}) = p_s\,].$$

In the simplest scenario, all subjects within a stratum would have exactly the same covariates, making **C** equal to zero. Intuitively, we can see that in this case we could not reduce the variance of the stratified treatment effect estimator by allowing further stratification on the covariates. Conversely, we might find that subjects within each stratum had the same propensity score but different covariate values, in which case we would expect **C** to be non-zero. Then if by chance in our sample dataset the covariate distributions happened to be imbalanced across treatment groups within the strata, we could reduce the variance of the estimator by further stratification on the covariates. This is exactly what stratification by the estimated propensity score would do here. In the more complicated situation where the strata contain more than one value of the propensity score, the term **C** is a measure of the average covariance between the outcomes and the covariates, taking account of the imbalance of the propensity score distributions across treatment groups within the strata.

We can therefore see that **C** measures the amount by which we can hope to reduce the variance of the stratified treatment effect estimator through creating strata based on the estimated propensity score which balance the covariates that are related to outcome better than they would be balanced by the strata based on the true propensity score.

3.4.4 The variance component V_4

From Theorem 3.2, the fourth variance component is

$$V_4 = \frac{1}{n}\,\mathbf{e}\,(n\,\mathrm{Cov}[\,\hat{\boldsymbol{\alpha}}\,])\,\mathbf{e}^T,$$

Recalling that m is the number of covariates used to estimate the propensity score, we define $\mathbf{e} = (e_1, e_2, ..., e_m)$ by $e_k = e_{\boldsymbol{\alpha}k} + e_{\mathbf{q}k}$ for $k = 1, ..., m$, where

$$e_{\boldsymbol{\alpha}k} = \sum_{s=1}^{K} r_s \left\{ \frac{(I_{Y_1 k} - \mathbb{E}[\,Y \mid Z = 1, S_{so} = 1\,]\, I_{f_1 k})}{d_{so}} - \frac{(I_{Y_0 k} - \mathbb{E}[\,Y \mid Z = 0, S_{so} = 1\,]\, I_{f_0 k})}{r_s - d_{so}} \right\},$$

and

$$I_{f_1 k} = \int_{q_{(s-1)o}}^{q_{so}} r \frac{\partial}{\partial \alpha_k} \{f_p(r; \boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\; dr$$

$$I_{f0k} = \int_{q_{(s-1)o}}^{q_{so}} (1-r)\frac{\partial}{\partial \alpha_k}\left\{f_p(r;\boldsymbol{\alpha})\right\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\, dr$$

$$I_{Y_1k} = \int_{q_{(s-1)o}}^{q_{so}} r\frac{\partial}{\partial \alpha_k}\left\{\mathbb{E}\left[\,Y_1 \,|\, p\left(\mathbf{X};\boldsymbol{\alpha}\right) = r\,\right] f_p(r;\boldsymbol{\alpha})\right\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\, dr$$

$$I_{Y_0k} = \int_{q_{(s-1)o}}^{q_{so}} (1-r)\frac{\partial}{\partial \alpha_k}\left\{\mathbb{E}\left[\,Y_0 \,|\, p\left(\mathbf{X};\boldsymbol{\alpha}\right) = r\,\right] f_p(r;\boldsymbol{\alpha})\right\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\, dr,$$

and

$$e_{\mathbf{q}k} = -\sum_{j=1}^{K-1} \left.\frac{\partial \beta^*}{\partial q_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} f_p(q_{jo})^{-1}\, \frac{\partial}{\partial \alpha_k}\left\{\mathbb{E}\left[\,1_{[p(\mathbf{X};\boldsymbol{\alpha})<q_j]}\,\right]\right\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

In order to understand what this component is measuring we first calculate the derivative of β^* with respect to α_k. The population parameter β^*, as mentioned before, is the same as the 'true' value of $\hat{\beta}^s$, β_o^s, but seen as a function of the parameters $\mathbf{q}$ and $\boldsymbol{\alpha}$, rather than evaluated at their population values. Therefore, the derivative of this parameter with respect to α_k measures the change in the population parameter being estimated, β_o^s, caused by a change in the propensity score parameter α_k. Keeping all other propensity score parameters fixed, the strata boundaries are functions of the propensity score parameter α_k. We can then view β^* as a function of the parameters $\{\alpha_k,\, q_1(\alpha_k),\, q_2(\alpha_k), \ldots,\, q_{K-1}(\alpha_k)\}$. Then the derivative of β^* with respect to α_k is

$$\left.\frac{d\beta^*}{d\alpha_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \left.\frac{\partial \beta^*}{\partial \alpha_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} + \sum_{j=1}^{K-1} \left.\frac{\partial \beta^*}{\partial q_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \left.\frac{\partial q_j}{\partial \alpha_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}. \tag{3.9}$$

Since the strata boundaries are estimated through the relationship

$$\mathbb{E}\left[\,1_{[p(\mathbf{X};\boldsymbol{\alpha})<q_j]}\,\right] = r_1 + r_2 + \ldots + r_j,$$

we can apply the usual rules of implicit differentiation to get,

$$\left.\frac{\partial q_j}{\partial \alpha_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -\frac{\left.\frac{\partial}{\partial \alpha_k}\left\{\mathbb{E}\left[\,1_{[p(\mathbf{X};\boldsymbol{\alpha})<q_j]}\,\right]\right\}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}}{\left.\frac{\partial}{\partial q_j}\left\{\mathbb{E}\left[\,1_{[p(\mathbf{X};\boldsymbol{\alpha})<q_j]}\,\right]\right\}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}} = -\frac{\left.\frac{\partial}{\partial \alpha_k}\left\{\mathbb{E}\left[\,1_{[p(\mathbf{X};\boldsymbol{\alpha})<q_j]}\,\right]\right\}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}}{f_p(q_{jo})}.$$

Then substituting this derivative into (3.9), we find that the derivative of β^* with respect to α_k is

$$\left.\frac{d\beta^*}{d\alpha_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \left.\frac{\partial\beta^*}{\partial\alpha_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} - \sum_{j=1}^{k-1} \left.\frac{\partial\beta^*}{\partial q_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} f_p(q_{oj})^{-1} \frac{\partial}{\partial\alpha_k} \left\{\mathbb{E}\left[1_{[p(\mathbf{X};\boldsymbol{\alpha})<q_j]}\right]\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

Now, we show in Appendix B.4.5 that $\left.\frac{\partial\beta^*}{\partial\alpha_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -C_k + e_{\boldsymbol{\alpha}k}$. Therefore,

$$\left.\frac{d\beta^*}{d\alpha_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -C_k + e_{\boldsymbol{\alpha}k} + e_{\mathbf{q}k}.$$

This last equation tells us that, keeping the other propensity score parameters fixed, changing the parameter α_k has two effects: (i) changing the balance of covariates associated with outcome across treatment groups within strata, measured by $-C_k$, and (ii) adding random error caused directly by the variability of the estimated propensity score parameters, measured by $e_{\boldsymbol{\alpha}k}$, and indirectly through adding to the variability of the estimated strata boundaries, measured by $e_{\mathbf{q}k}$.

The variance component V_3 has already taken into consideration the reduction in variance of the stratified treatment effect estimator caused by the estimation of the propensity score, through increasing the covariate balance within strata. The last variance component, V_4, measures the increase in variance caused by the random error introduced by the estimation of the propensity score. The component V_4 is a quadratic form around a positive-definite matrix and so is always positive. The total effect of the estimation of the propensity score on the variance of the stratified treatment effect estimator is measured by the sum $V_3 + V_4$. If we could show theoretically that the magnitude of V_4 is always smaller than the magnitude of V_3 then we would know that estimation of the propensity score always decreases the variance. This, however, has not yet been possible due to the complex nature of V_4. In the simulated examples that we consider later (Section 5.2) it is always true that $V_3 + V_4 < 0$, indicating that in these examples the estimation of the propensity score decreases the variance of the stratified treatment effect estimator. In some cases, however, the magnitudes of the two components are almost equal.

3.5 Discussion

In this chapter we have ascertained the theoretical properties of the stratified treatment effect estimator, $\hat{\beta}^s$, in the situations where the propensity score is: (i) a known

function of the observed covariates, and (ii) estimated using a correctly specified logistic regression model. In particular, under each of these assumptions, we derived conditions under which $\hat{\beta}^s$ is consistent and asymptotically normally distributed, and we calculated its asymptotic variance. The two asymptotic variances are denoted by $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$ for the situations where the propensity score is known and estimated, respectively. We further showed that these two variance formulæ could be expressed in terms of four variance components as follows.

$$\mathbb{V}_k[\hat{\beta}^s] = V_1 + V_2$$
$$\mathbb{V}_e[\hat{\beta}^s] = V_1 + V_2 + V_3 + V_4,$$

where

$V_1 =$ The variance when the population strata boundaries and propensity score are known.

$V_2 =$ The increase in variance due to the random error introduced by estimating the strata boundaries.

$V_3 =$ The reduction in variance due to the increased balance of covariates associated with the outcome within strata, caused by estimating the propensity score.

$V_4 =$ The increase in variance due to the random error introduced by estimating the propensity score.

Conditions under which we expect these variance formulæ to be valid were derived and discussed. In particular, the conditions which required both the probability density function of the propensity score and the conditional expectation of the outcome given treatment status and propensity score, to be bounded and differentiable, are the conditions most likely to be violated, without the violation being apparent during the analysis of the dataset. We can imagine practical situations where these conditions are likely to be almost violated, in which case we might require very large sample sizes in order for the two variance formulæ given above to be valid. These issues are further investigated in Chapter 5.

We have seen, in this chapter, that the difficulty in finding a suitable variance estimator for $\hat{\beta}^s$ comes from the non-continuity of the strata indicators under repeated

sampling from the near-infinite population. However, we can easily calculate the conditional variance of $\hat{\beta}^s$, viewing our covariates and treatment status as fixed, since the strata indicators are then fixed and therefore the problem of non-continuity vanishes. We discuss, in Chapter 4, whether such a conditional variance estimator or a marginal estimator — where the covariates and treatment status indicators are seen as random variables under repeated sampling — is preferable. We show that our variance estimators, $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$, measure the asymptotic marginal variance of β^s.

Having calculated the variances $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$ and explored them mathematically, we then apply them to hypothetical situations in Chapter 5. We take several simple example situations with known outcome, covariate and propensity score distributions, allowing us to calculate the 'true' values of $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$, without having to estimate them from a sample dataset. We then obtain empirical variance estimators, by simulating many sample datasets from each example, in order to estimate the sampling distribution of $\hat{\beta}^s$. We compare the theoretical and empirical estimates of $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$, in order to investigate how large a sample is needed in order for the empirical variances to be close to the theoretical asymptotic variances. We also consider two examples where one of the conditions required for the validity of $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$ is almost violated, in order to see how this affects the convergence rate.

We then proceed, in Chapter 6, to consider the issue of estimating the two variance estimators from a sample dataset. Whilst the component V_1 can be easily estimated from a sample dataset, and V_2 can be expected to be negligible, and therefore ignored, neither of the components V_3 and V_4 can be expected to be negligible and the component V_4, in particular, is not easy to estimate from a dataset.

An alternative approach to calculating the variance is as follows. Rather than multiplying out $A^{-1}BA^{-T}$ and calculating an explicit formula for the variance, we could merely replace the expectations in A and B by sample averages and hence estimate $A^{-1}BA^{-T}$ by multiplying the sample estimates of the two matrices. However, the presence of intractable derivatives in the matrix A complicates this approach. We will, however, return later to this idea.

Chapter 4

The marginal and conditional variances of the stratified treatment effect estimator

We now digress a little and discuss two ways of measuring the uncertainty in an estimate of a statistic: the marginal and conditional variances. Which of these types of variance is more appropriate is determined by the parameter of interest. We may wish to estimate a causal parameter that relates only to the sample at hand, in which case the conditional variance, given the characteristics of this sample, would be the appropriate measure of variance. In our problem, this is the variance of the stratified treatment effect estimator conditional on the observed distribution of the treatment status and covariates. Conversely, we might be interested in the parameters of the near-infinite population from which the data were sampled, a premise which we assumed to be true in the previous chapters. In this situation, the marginal variance, which views all characteristics of the sample as random variables, would be the appropriate measure of variance. We begin this chapter with a brief discussion of the merits of both types of variance, with emphasis on our particular situation. We then calculate the conditional variance of the stratified treatment effect estimator, conditional on treatment status and covariates. Assuming that the propensity score is a known function of the observed covariates, we then marginalise this conditional variance over the distribution of the treatment status and covariates. In this calculation we use first-order approximations and so the resulting variance is a first-order approximation to the marginal variance of the stratified treatment effect estimator. This turns out to be the variance calculated in Chapter 3, $\mathbb{V}_k[\,\hat{\beta}^s\,]$. In this way, we show that $\mathbb{V}_k[\,\hat{\beta}^s\,]$, and by extension, $\mathbb{V}_e[\,\hat{\beta}^s\,]$, are asymptotic (first-order) marginal variances of the stratified treatment effect estimator.

4.1 The relationship between marginal and conditional variances

We now define the conditional and marginal variances mathematically and describe the relationship between the two. The marginal variance considers all observed variables to be random variables. The conditional variance, on the other hand, treats all observed covariates as fixed quantities and allows only the outcome to vary, given that fixed covariate structure. Standard theory [7, p.154] shows that if we have random variables Y and X with $Y = Y(X)$ then the marginal variance of Y is

$$\mathbb{V}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2],$$

and the conditional variance of Y given X is

$$\mathbb{V}[Y \mid X] = \mathbb{E}[(Y - \mathbb{E}[Y \mid X])^2 \mid X].$$

These two variances are linked by the formula

$$\mathbb{V}[Y] = \mathbb{E}_X[\mathbb{V}[Y \mid X]] + \mathbb{V}_X[\mathbb{E}[Y \mid X]]. \tag{4.1}$$

The second term in (4.1) cannot be negative since it is a variance. Therefore, the marginal variance of Y must be at least as large as the expectation of the conditional variances. Typically, a marginal variance estimate will be larger than a conditional variance estimate.

4.1.1 Are marginal or conditional variances more appropriate?

Although some epidemiologists believe that we should only attempt to estimate causal parameters for the sample at hand [69], we often wish to generalise our results to a wider population. For example, in public health, we may wish to estimate the treatment effect we would see if we chose to make an intervention available throughout the country, in which case the most appropriate variance would be the marginal variance.

There are many examples of both marginal and conditional variances used routinely in epidemiology. Methods such as general estimating equations produce marginal estimates and variances, and robust standard errors are marginal. By contrast, in a standard regression model we assume that the covariates and treatment indicators

are fixed and so both the treatment effect estimate and the variance of the treatment effect estimate are conditional on these covariates and treatment indicators. The Mantel-Haenszel odds ratio also conditions on the fixed strata.

In our situation, viewing the strata as fixed variables would remove the problems of non-continuity, greatly simplifying the variance formulæ. There is, however, one key difference between stratifying on the propensity score and stratifying on, for example, age. If we calculated a Mantel-Haenszel odds ratio for a particular treatment and outcome, and wished to stratify by age, we might reasonably group subjects into 5- or 10-year categories and then treat these strata as fixed. These age categories then have a clinical meaning, as do the resulting within-strata odds ratios. The propensity score categories, however, have no such clinical meaning. It therefore seems appropriate to marginalise over the covariate and treatment status distributions.

4.2 The marginal variance of the stratified treatment effect estimator

We now calculate the variance of the stratified treatment effect estimator conditional on the observed treatment status indicators and covariates. We then assume that the propensity score is a known function of the observed covariates and marginalise the conditional variance over the distribution of the treatment status indicators and covariates using (4.1). In this way we obtain the marginal variance of the stratified treatment effect estimator.

4.2.1 The conditional variance given treatment and covariates

In the notation introduced in Section 3.1.1 when the propensity score is a known function of the observed covariates the stratified treatment effect estimator is defined as

$$\hat{\beta}^s = \frac{1}{n}\sum_{s=1}^{K} r_s \sum_{i=1}^{n}\left\{\frac{Y_i\, Z_i\, \hat{S}_{si}}{\hat{d}_s} - \frac{Y_i\,(1-Z_i)\,\hat{S}_{si}}{r_s - \hat{d}_s},\right\},$$

where $\hat{S}_{si} = 1_{[\hat{q}_{(s-1)} < p_o(\mathbf{X}_i) < \hat{q}_s]}$ is an indicator for the s^{th} sample stratum and $p_o(\mathbf{X}_i)$ is the known propensity score. This can be re-written as

$$\hat{\beta}^s = \sum_{s=1}^{K} r_s \sum_{i=1}^{n}\left\{\frac{Y_i\, Z_i\, \hat{S}_{si}}{\sum_{i=1}^{n} Z_i\, \hat{S}_{si}} - \frac{Y_i\,(1-Z_i)\,\hat{S}_{si}}{\sum_{i=1}^{n}(1-Z_i)\,\hat{S}_{si}}\right\}.$$

Assuming that the treatment status and observed covariates for each sampled subject are fixed, the outcomes, Y_i, are the only quantities which can vary. Further assuming that subjects are independent and identically distributed gives the conditional variance of $\hat{\beta}^s$,

$$\mathbb{V}\left[\hat{\beta}^s \mid Z, \mathbf{X}\right] = \sum_{s=1}^{K} r_s^2 \, (\zeta_1 + \zeta_2), \tag{4.2}$$

where

$$\zeta_1 = \frac{\sum_{i=1}^{n} Z_i \, \hat{S}_{si} \, \mathbb{V}\left[Y \mid Z = 1, \mathbf{X} = X_i\right]}{\left(\sum_{i=1}^{n} Z_i \, \hat{S}_{si}\right)^2}$$

$$\zeta_2 = \frac{\sum_{i=1}^{n} (1 - Z_i) \, \hat{S}_{si} \, \mathbb{V}\left[Y \mid Z = 0, \mathbf{X} = X_i\right]}{\left(\sum_{i=1}^{n} (1 - Z_i) \, \hat{S}_{si}\right)^2}$$

Note that if we wished to use this variance in practice, we would have to fit some sort of model to estimate the variance of the outcome conditional on the treatment status and covariates.

4.2.2 Marginalising the conditional variance

Assuming that the propensity score is a known function of the observed covariates, we can obtain the marginal variance of the stratified treatment effect estimator from the conditional variance formula (4.2) using the general equation linking marginal and conditional variances (4.1),

$$\mathbb{V}\left[\hat{\beta}^s\right] = \mathbb{E}_{Z,\mathbf{X}}\left[\mathbb{V}\left[\hat{\beta}^s \mid Z, \mathbf{X}\right]\right] + \mathbb{V}_{Z,\mathbf{X}}\left[\mathbb{E}\left[\hat{\beta}^s \mid Z, \mathbf{X}\right]\right].$$

The following lemma greatly simplifies the process of marginalising the conditional variance.

Lemma 4.1

$$\begin{aligned} &\mathbb{E}_{Z,\mathbf{X}}\left[\mathbb{V}\left[\hat{\beta}^s \mid Z, \mathbf{X}\right]\right] + \mathbb{V}_{Z,\mathbf{X}}\left[\mathbb{E}\left[\hat{\beta}^s \mid Z, \mathbf{X}\right]\right] \\ = \; &\mathbb{E}_{Z,\hat{\mathbf{S}}}\left[\mathbb{V}\left[\hat{\beta}^s \mid Z, \hat{\mathbf{S}}\right]\right] + \mathbb{V}_{Z,\hat{\mathbf{S}}}\left[\mathbb{E}\left[\hat{\beta}^s \mid Z, \hat{\mathbf{S}}\right]\right]. \end{aligned} \tag{4.3}$$

Proof

By applying the general relationship between marginal and conditional variances, we can write

$$\begin{aligned}\mathbb{E}_{Z,\hat{\mathbf{S}}}[\mathbb{V}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}}]] &= \mathbb{E}_{Z,\hat{\mathbf{S}}}[\mathbb{E}_{\mathbf{X}}[\mathbb{V}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}},\mathbf{X}]]] + \mathbb{E}_{Z,\hat{\mathbf{S}}}[\mathbb{V}_{\mathbf{X}}[\mathbb{E}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}},\mathbf{X}]]] \\ &= \mathbb{E}_{Z,\mathbf{X}}[\mathbb{V}[\hat{\beta}^s \mid Z,\mathbf{X}]] + \mathbb{E}_{Z,\hat{\mathbf{S}}}[\mathbb{V}_{\mathbf{X}}[\mathbb{E}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}},\mathbf{X}]]], \qquad (4.4)\end{aligned}$$

where the second line follows since $\mathbf{X}$ completely determines $\hat{S}$. By the usual rules of iterated expectation, we also have

$$\mathbb{V}_{Z,\hat{\mathbf{S}}}[\mathbb{E}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}}]] = \mathbb{V}_{Z,\hat{\mathbf{S}}}[\mathbb{E}_{\mathbf{X}}[\mathbb{E}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}},\mathbf{X}]]]. \qquad (4.5)$$

Therefore, combining (4.4) and (4.5) gives

$$\begin{aligned}&\mathbb{E}_{Z,\hat{\mathbf{S}}}[\mathbb{V}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}}]] + \mathbb{V}_{Z,\hat{\mathbf{S}}}[\mathbb{E}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}}]] \\ &= \mathbb{E}_{Z,\mathbf{X}}[\mathbb{V}[\hat{\beta}^s \mid Z,\mathbf{X}]] \\ &+ \mathbb{E}_{Z,\hat{\mathbf{S}}}[\mathbb{V}_{\mathbf{X}}[\mathbb{E}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}},\mathbf{X}]]] + \mathbb{V}_{Z,\hat{\mathbf{S}}}[\mathbb{E}_{\mathbf{X}}[\mathbb{E}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}},\mathbf{X}]]]. \qquad (4.6)\end{aligned}$$

A final application of the general relationship between marginal and conditional variances yields

$$\mathbb{E}_{Z,\hat{\mathbf{S}}}[\mathbb{V}_{\mathbf{X}}[\mathbb{E}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}},\mathbf{X}]]] + \mathbb{V}_{Z,\hat{\mathbf{S}}}[\mathbb{E}_{\mathbf{X}}[\mathbb{E}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}},\mathbf{X}]]] = \mathbb{V}_{Z,\mathbf{X}}[\mathbb{E}[\hat{\beta}^s \mid Z,\mathbf{X}]].$$

And substituting this equality into (4.6) gives, as required

$$\begin{aligned}&\mathbb{E}_{Z,\hat{\mathbf{S}}}[\mathbb{V}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}}]] + \mathbb{V}_{Z,\hat{\mathbf{S}}}[\mathbb{E}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}}]] \\ &= \mathbb{E}_{Z,\mathbf{X}}[\mathbb{V}[\hat{\beta}^s \mid Z,\mathbf{X}]] + \mathbb{V}_{Z,\mathbf{X}}[\mathbb{E}[\hat{\beta}^s \mid Z,\mathbf{X}]].\end{aligned}$$

◇

Therefore, assuming that the propensity score is a known function of the observed covariates we can calculate the marginal variance of the stratified treatment effect estimator from the conditional variance (4.2) using the following relationship,

$$\mathbb{V}[\hat{\beta}^s] = \mathbb{E}_{Z,\hat{\mathbf{S}}}[\mathbb{V}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}}]] + \mathbb{V}_{Z,\hat{\mathbf{S}}}[\mathbb{E}[\hat{\beta}^s \mid Z,\hat{\mathbf{S}}]]. \qquad (4.7)$$

We now proceed to calculate the expectation and variance above.

Calculating $\mathbb{V}_{\mathbf{Z},\hat{\mathbf{S}}}[\mathbb{E}[\hat{\beta}^s \mid \mathbf{Z}, \hat{\mathbf{S}}]]$

Taking the conditional expectation of $\hat{\beta}^s$ given treatment status, Z, and estimated strata, $\hat{\mathbf{S}}$, gives

$$\mathbb{E}[\hat{\beta}^s \mid Z, \hat{\mathbf{S}}] = \sum_{s=1}^{K} r_s \left\{ \sum_{i=1}^{n} \frac{Z_i \hat{S}_{si} \mathbb{E}[Y \mid Z = 1, \hat{S}_s = 1]}{\sum_{i=1}^{n} Z_i \hat{S}_{si}} - \sum_{i=1}^{n} \frac{(1 - Z_i) \hat{S}_{si} \mathbb{E}[Y \mid Z = 0, \hat{S}_s = 1]}{\sum_{i=1}^{n} (1 - Z_i) \hat{S}_{si}} \right\}.$$

This cancels to give

$$\mathbb{E}[\hat{\beta}^s \mid Z, \hat{\mathbf{S}}] = \sum_{s=1}^{K} r_s \{\mathbb{E}[Y \mid Z = 1, \hat{S}_s = 1] - \mathbb{E}[Y \mid Z = 0, \hat{S}_s = 1]\}.$$

We now wish to calculate the variance of this conditional expectation over the joint distribution of the treatment status and estimated strata, Z and $\hat{\mathbf{S}}$. Since the indicator $\hat{S}_s = 1_{[\hat{q}_{s-1} \le p_o(\mathbf{X}) < \hat{q}]}$ depends only on the distribution of the random variable $p_o(\mathbf{X})$ and the estimated strata boundaries, the expectations $\mathbb{E}[Y \mid Z = t, \hat{S}_s = 1]$, for $t = 0, 1$, depend on the observed Z_i and $\hat{S}_{si}$ only through the estimated strata boundaries, $\hat{\mathbf{q}}$. Therefore,

$$\mathbb{V}_{Z,\hat{\mathbf{S}}}[\mathbb{E}[\hat{\beta}^s \mid Z, \hat{\mathbf{S}}]] = \mathbb{V}_{\hat{\mathbf{q}}}[\mathbb{E}[\hat{\beta}^s \mid Z, \hat{\mathbf{S}}]]. \tag{4.8}$$

We can view $\mathbb{E}[\hat{\beta}^s \mid Z, \hat{\mathbf{S}}]$ as a function of $\hat{\mathbf{q}}$ and write $\mathbb{E}[\hat{\beta}^s \mid Z, \hat{\mathbf{S}}] = \beta^*(\hat{\mathbf{q}})$. A multivariate Taylor series expansion of $\beta^*(\hat{\mathbf{q}})$ about $\mathbf{q}_o$ gives

$$\mathbb{E}[\hat{\beta}^s \mid Z, \hat{\mathbf{S}}] = \beta^*(\hat{\mathbf{q}}) = \beta^*(\mathbf{q}_o) + \left.\frac{\partial \beta^*}{\partial \mathbf{q}^T}\right|_{\mathbf{q}=\mathbf{q}_o} (\hat{\mathbf{q}} - \mathbf{q}_o) + O_p((\hat{\mathbf{q}} - \mathbf{q}_o)^2).$$

Ignoring terms of order less than or equal to $(\hat{\mathbf{q}} - \mathbf{q}_o)^2$, we take the variance of the above equation. Using (4.8) we then see that the required variance is asymptotically equal to

$$\mathbb{V}_{Z,\hat{\mathbf{S}}}[\mathbb{E}[\hat{\beta}^s \mid Z, \hat{\mathbf{S}}]] = \left.\frac{\partial \beta^*}{\partial \mathbf{q}^T}\right|_{\mathbf{q}=\mathbf{q}_o} \mathrm{Cov}[\hat{\mathbf{q}}] \left.\frac{\partial \beta^*}{\partial \mathbf{q}}\right|_{\mathbf{q}=\mathbf{q}_o}.$$

This is equal to the variance component V_2 (see Theorem 3.1).

Calculating $\mathbb{E}_{\mathbf{Z},\hat{\mathbf{S}}}[\mathbb{V}[\hat{\beta}^s \mid \mathbf{Z}, \hat{\mathbf{S}}]]$

Taking the conditional variance of $\hat{\beta}^s$ given treatment status, Z, and estimated strata, $\hat{\mathbf{S}}$, gives

$$\mathbb{V}[\hat{\beta}^s \mid Z, \hat{\mathbf{S}}] = \sum_{s=1}^{K} r_s^2 \left\{ \frac{\mathbb{V}[Y \mid Z=1, \hat{S}_s=1]}{\sum_{i=1}^{n} Z_i \hat{S}_{si}} + \frac{\mathbb{V}[Y \mid Z=0, \hat{S}_s=1]}{\sum_{i=1}^{n}(1-Z_i)\hat{S}_{si}} \right\}.$$

A multivariate Taylor series expansion of the above variance about $\mathbf{q}_o$ ignoring terms of order less than or equal to $(\hat{\mathbf{q}}-\mathbf{q}_o)^2$, and using first-order approximations [12, p.79] shows that the required expectation is asymptotically equal to

$$\mathbb{E}_{Z,\hat{\mathbf{S}}}[\mathbb{V}[\hat{\beta}^s \mid Z, \hat{\mathbf{S}}]] = \sum_{s=1}^{K} \frac{r_s^2}{n} \left\{ \frac{\mathbb{V}[Y \mid Z=1, S_{so}=1]}{d_{so}} + \frac{\mathbb{V}[Y \mid Z=0, S_{so}=1]}{r_s - d_{so}} \right\}.$$

This is equal to the variance component V_1 (see Theorem 3.1).

Using (4.7) we see that the marginal variance of the stratified treatment effect estimator, when the propensity score is a known function of the observed covariates, is

$$\mathbb{V}[\hat{\beta}^s] = V_1 + V_2 + O_p((\hat{\mathbf{q}} - \mathbf{q})^2).$$

This demonstrates that the variance $\mathbb{V}_k[\hat{\beta}^s] = V_1 + V_2$ that we calculated in the previous chapter is a first-order approximation to the marginal variance of the stratified treatment effect estimator, when the propensity score is a known function of the observed covariates. By extension, we conjecture that the variance $\mathbb{V}_e[\hat{\beta}^s]$ is also a first-order approximation to the marginal variance of the stratified treatment effect estimator, when the propensity score is estimated using a correctly specified logistic regression model.

4.3 The variance formula used in applications

Lunceford and Davidian [59] state that the routine variance formula for the stratified treatment effect estimator used in applications is

$$\mathbb{V}[\hat{\beta}^s] = \sum_{i=1}^{N} \left(\frac{n_s}{n}\right)^2 \left\{ \frac{(\hat{\sigma}_s^t)^2}{n_s^t} + \frac{(\hat{\sigma}_s^c)^2}{n_s^c} \right\}, \tag{4.9}$$

where $\hat{\sigma}_s^t$ is a sample estimate of the variance of the outcome amongst treated subjects in the s^{th} sample stratum, n_s denotes the number of subjects in the s^{th} sample

stratum, n_s^t denotes the number of treated subjects in the s^{th} sample stratum, and the quantities with a superscript of 'c' are defined similarly for the untreated group.

This is a sample estimate of the variance component V_1. We have already noted that we expect the variance component V_2 to be negligible. Therefore, (4.9) is essentially a sample estimate of $\mathbb{V}_k[\,\hat{\beta}^s\,]$, the asymptotic marginal variance of the stratified treatment effect estimator when the propensity score is a known function of the observed covariates.

4.4 Discussion

In this chapter we calculated the variance of the stratified treatment effect estimator, conditional on treatment status and covariates. We condition in this way — treat the covariates and treatment status indicators as fixed — in a standard regression model. We then showed that by marginalising this conditional variance over the distribution of the covariates and treatment status indicators, assuming that the propensity score is a known function of the observed covariates, and using first-order approximations, we obtain $\mathbb{V}_k[\,\hat{\beta}^s\,]$, the variance calculated in Chapter 3. In this way, we see that $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$ are asymptotic marginal variances of the stratified treatment effect estimator.

We can visualise the difference between the conditional variance and the marginal variance as follows. If we repeatedly sampled datasets of n subjects from our near-infinite population, each time selecting $\{Y_i, Z_i, \mathbf{X}_i\}$ afresh, and calculated the stratified treatment effect estimator for each dataset, then the variance of these treatment effect estimators would be the marginal variance — either $\mathbb{V}_k[\,\hat{\beta}^s\,]$ or $\mathbb{V}_e[\,\hat{\beta}^s\,]$, depending on whether we had used a known or estimated propensity score. If, on the other hand, we kept the covariates and treatment status indicators from the first sample and merely re-sampled the outcomes, given the observed treatment status indicators and covariates, then the variance among the resulting stratified treatment effect estimators would be the conditional variance. Note that the distinction between marginal and conditional variances here is concerned with whether the covariates and treatment status indicators are viewed as fixed or random, and does not refer to marginalising over the estimated propensity score parameters.

We showed that the variance formula for the stratified treatment effect estimator commonly used in applications is essentially a sample estimate of $\mathbb{V}_k[\hat{\beta}^s]$, the asymptotic marginal variance when the propensity score is a known function of the observed covariates. Since the propensity score is invariably unknown and must be estimated from the data we should use $\mathbb{V}_e[\hat{\beta}^s]$. In the following chapter, we will see that in all the hypothetical examples we consider, $\mathbb{V}_e[\hat{\beta}^s]$ is smaller than $\mathbb{V}_k[\hat{\beta}^s]$, often substantially smaller. This suggests that the variance estimator being used in practice is overestimating the variance and thus the resulting confidence intervals and hypothesis tests are too conservative.

We now return to the two variances that we calculated in the previous chapter, $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$, that we have now shown to be asymptotic marginal variances. In Chapter 5 we apply these variances to various hypothetical situations and look at how large a sample is needed in order for these asymptotic variances to be valid.

Chapter 5

Practical performance of the variance formulæ for the stratified treatment effect estimator

In Chapter 3, the variance of the stratified treatment effect estimator was calculated, assuming that the propensity score is: (i) a known function of the observed covariates, and (ii) estimated from the data using a correctly specified logistic regression model. We expressed these two variances, denoted by $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$, respectively, in terms of four variance components, V_1, V_2, V_3 and V_4 (Theorems 3.1 and 3.2). In Section 3.4 we discussed the source of error measured by each of these variance components. We therefore begin this chapter by calculating the four variance components for a simple hypothetical example, varying each of the parameters of the example one at a time in order to see if the change in variance components accords with our intuition, gained through this discussion.

We then proceed to investigate the convergence rates of the asymptotic variances $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$. Theorems 3.1 and 3.2 give conditions under which we can expect the stratified treatment effect estimator to be asymptotically normally distributed, with asymptotic variance equal to $\mathbb{V}_k[\,\hat{\beta}^s\,]$ or $\mathbb{V}_e[\,\hat{\beta}^s\,]$. Provided that these conditions are satisfied, the variance formulæ should be valid for 'large enough' samples. How large is large enough will depend on how close the example comes to violating the conditions. We therefore consider four simple hypothetical situations. The first two do not violate any of the conditions but each of the remaining two almost violate one of the conditions. For each of the four hypothetical examples, we calculate $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$ using the formulæ given in Theorems 3.1 and 3.2. Since we know the distribution of the data we calculate the two variances mathematically, without having to estimate them from a simulated dataset. We then use simulation to obtain

empirical estimates of the two variances, for various sample sizes. This gives us some indication of how large a sample size is needed for these two variance formulæ to be valid in practice.

The four variance components all tend to zero as the sample size gets large. Each variance component is a function of the form k/n where n is the sample size and k is a constant that does not depend on the sample size. For this reason, in this chapter we discuss the behaviour of $n\,V_1, n\,V_2, n\,V_3$ and $n\,V_4$, or $n\,\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $n\,\mathbb{V}_e[\,\hat{\beta}^s\,]$, since these are constants which do not depend on the sample size chosen.

5.1 Application of the variance formulæ to a hypothetical example

We now introduce a simple hypothetical example, and calculate the four variance components for this example. Since we specify the distribution of the data, it is possible to calculate these variance components mathematically, without estimating them from a dataset. This calculation is described briefly below. We vary the parameters of the example one at a time to see whether the resulting changes in the variance components are in the anticipated direction, given the discussion of these variance components in Section 3.4.

5.1.1 A hypothetical example

Our hypothetical example has two covariates – a binary covariate, X_1, and a continuous covariate, X_2, with a distribution that depends on X_1. The propensity score depends on both covariates. The outcome, Y, is continuous and depends on the covariate X_2 and treatment status, Z, only. The individual-level treatment effect is the same for all subjects and is therefore equal to the population average causal treatment effect. We define five population strata, each of which contains an equal fraction of the whole population. Details are as follows:

$$\begin{aligned}
\text{Outcome:} \quad & Y = \gamma_0 + \gamma_2\, X_2 + 2\, Z + \epsilon, \quad \epsilon \sim N(0, 10^2). \\
\text{Propensity score:} \quad & \ln\left(\frac{p_o(\mathbf{X})}{1 - p_o(\mathbf{X})}\right) = \alpha_0 + \alpha_1\, X_1 + \alpha_2\, X_2. \qquad (5.1) \\
\text{Covariates:} \quad & \mathbb{P}(X_1 = 0) = 0.6, \\
& X_2 \,|\, X_1 = 0 \sim N(70, 10^2), \quad X_2 \,|\, X_1 = 1 \sim N(60, 15^2).
\end{aligned}$$

The 'default' values are: $(\gamma_0, \gamma_2) = (35, -0.35)$, and $(\alpha_0, \alpha_1, \alpha_2) = (-2.0.15.0.01)$. Each of these parameter values are varied one at a time, while the others remain fixed at their default value.

The calculation of the variance components

The four variance components, V_1, V_2, V_3 and V_4, are defined in Theorems 3.1 and 3.2. To calculate the 'true' values of these four variance components in the hypothetical example (5.1), it is necessary to calculate various expectations, probabilities and other population quantities. We give a brief outline of the calculation here. Full details of the calculation can be found in Appendix C.

First, the probability density function of the propensity score in model (5.1) is calculated using the Jacobian change of variables method, giving

$$f_p(p) = \frac{0.6\, v_0(p) + 0.4\, v_1(p)}{|\alpha_2|\, p\,(1-p)}, \tag{5.2}$$

where

$$v_0(p) = \exp\left\{-\frac{(w_0(p) - 70)^2}{2 \times 10^2}\right\} \Big/ \sqrt{2\pi 10^2} \qquad w_0(p) = \frac{\ln\left(\frac{p}{1-p}\right) - \alpha_0}{\alpha_2},$$

$$v_1(p) = \exp\left\{-\frac{(w_1(p) - 60)^2}{2 \times 15^2}\right\} \Big/ \sqrt{2\pi 15^2} \qquad w_1(p) = \frac{\ln\left(\frac{p}{1-p}\right) - \alpha_0 - \alpha_1}{\alpha_2}.$$

All probabilities and expectations involved in the four variance components are expressed as integrals over this distribution of the propensity score. The integrals are then calculated by approximate numerical methods using the mathematical software *Mathematica* [112]. We show how the first population strata boundary, q_{1o}, and the population probability of being treated and in the first stratum, d_{1o}, are calculated. The other quantities involved in the variance components are calculated in a similar way.

Since we have five equal sized strata, the first population strata boundary solves the equation $\mathbb{P}(p_o(\mathbf{X}) < q_{1o}) = 1/5$. We therefore find q_{1o} by solving the equation

$$\int_0^{q_{1o}} f_p\,(p)\; dp = 1/5.$$

Then q_{1o} is found by numerical approximation methods using the function **FindRoot** in the software *Mathematica*.

Remembering that $S_{1o} = 1_{[0 \leq p_o(\mathbf{X}) < q_{1o}]}$ is an indicator for the first population stratum, the population probability of being treated and in the first stratum, d_{1o}. is defined as

$$d_{1o} = \mathbb{P}(Z = 1,\, S_{1o} = 1) = \mathbb{E}\,[\,Z\,S_{1o}\,] = \mathbb{E}\,[\,p_o(\mathbf{X})\,S_{1o}\,].$$

We can therefore write the parameter d_{1o} in integral form as

$$d_{1o} = \int_0^{q_{1o}} p\; f_p\,(p)\; dp.$$

Using the value for q_{1o} calculated above, we calculate d_{1o} using the numerical integration function **NIntegrate** in the software *Mathematica*.

All remaining quantities contained in the four variance components are calculated in a similar way. The Mathematica program that calculates these variance components can be found in Appendix D.2.

5.1.2 Change in variance components as the outcome parameters vary

We begin by investigating the effect of varying the parameters of the outcome, γ_0 and γ_2, on the four variance components.

Varying the parameter γ_0

Changing the parameter γ_0 merely changes the outcome of each subject by the same amount. Therefore, the variance of the stratified treatment effect estimator should not depend on the value of γ_0. Three different values of γ_0 were tried: -35, 4 and 35. The resulting variance components are shown in Table 5.1. As expected, the value of γ_0 did not affect any of these components.

Note that in Table 5.1 the first variance component accounts for the majority of the variance and the second variance component, as expected, is negligible. The third variance component is not negligible and indicates that strata based on the estimated propensity score would be more balanced in the covariate X_2 — the only covariate that affects the outcome — than strata based on the true propensity score. The

Table 5.1: *Change in variance components as γ_0 varies.*

γ_0	-35	4	35
$n\,V_1$	637.82	637.82	637.82
$n\,V_2$	0.000008	0.000008	0.000008
$n\,V_3$	-36.82	-36.82	-36.82
$n\,V_4$	1.69	1.69	1.69

fourth variance component, as expected, is small in comparison to the first variance component, but much larger than the second, indicating that the estimation of the propensity score introduces more random error into the stratified treatment effect estimator than the estimation of the strata boundaries does.

Varying the parameter γ_2

The parameter γ_2 measures the effect of the covariate X_2 on the outcome. When $\gamma_2 = 0$, all variance of the outcome within a stratum and treatment group is due to random variation. As the magnitude of the covariate effect on the outcome increases, the variance of the outcome within the strata and treatment groups should increase. This part of the variance is measured by the variance component V_1. Therefore, we expect the variance component V_1 to be at its smallest when $\gamma_2 = 0$ and to increase with the magnitude of γ_2.

In the absence of any covariate effect on the outcome, estimation of the strata boundaries should not affect the variance of the stratified treatment effect estimator. Therefore, we expect $V_2 = 0$ when $\gamma_2 = 0$. As the magnitude of the covariate effect increases, the error due to estimating the strata boundaries should increase and so we expect V_2 to increase with the magnitude of γ_2.

With no covariate effect on the outcome, the estimation of the propensity score should have no effect on the variance of the stratified treatment effect estimator. Therefore, we expect $V_3 = 0$ when $\gamma_2 = 0$. As the magnitude of the covariate effect increases, the covariance between the outcome and the covariate X_2 within the population strata will increase and so the magnitude of the variance component V_3 should increase with the magnitude of γ_2. This indicates that as the magnitude of the covariate effect increases the potential for increasing the within-strata balance through using the estimated rather than the true propensity score also increases.

Since the estimation of the propensity score should have no effect on the variance of the stratified estimator of treatment effect when $\gamma_2 = 0$, we expect $V_4 = 0$. As the covariate effect on the outcome increases, the random error due to estimation of the propensity score should increase. Therefore, we expect V_4 to increase with the magnitude of γ_2.

Four values of γ_2 are tried: -5, -0.35, 0 and 5. All other parameters are set to their default values. The changes in variance components as γ_2 varies can be seen in Table 5.2. All changes are in the directions predicted.

Table 5.2: *Change in variance components as γ_2 varies.*

γ_2	-5	-0.35	0	5
$n\,V_1$	10361.88	637.82	589.94	10361.88
$n\,V_2$	0.0017	0.000008	0	0.0017
$n\,V_3$	-7513.04	-36.82	-1.49^{-26}	-7514.04
$n\,V_4$	344.16	1.69	1.29^{-25}	344.16

Again, V_1 accounts for most of the variance and V_2 is negligible. When $\gamma_2 = -5$, the magnitude of the component V_3 is about 70% of V_1. Therefore, in this situation, a substantial proportion of the variance can be removed by the estimation of the propensity score. As before, V_4 is small in comparison to V_1 but not always negligible.

5.1.3 Change in variance components as the propensity score parameters vary

We now investigate the effect of varying the parameters of the propensity score, α_1 and α_2, on the four variance components.

Varying the parameter α_1

The parameter α_1 measures the extent to which the covariate X_1 affects the propensity score. Since the outcome depends on the covariate X_2 and treatment status only, the smallest outcome variance within the strata and treatment groups would be achieved by stratifying on X_2 only. When $\alpha_1 = 0$, stratifying by the propensity score is exactly equivalent to stratifying by the covariate X_2. As the magnitude of α_1 increases, the variance of the covariate X_2 within strata and treatment groups will increase and therefore the variance of the outcome within strata and treatment groups will increase. So we expect the variance component V_1 to increase with the magnitude

of α_1. Note that since the propensity score is not a symmetric function of α_1, the component V_1 may not increase linearly with the magnitude of α_1.

We have noted that when $\alpha_1 = 0$ the smallest outcome variance within the strata and treatment groups is achieved by stratifying on the true propensity score. Therefore, we do not expect the estimation of the propensity score, when $\alpha_1 = 0$, to reduce the variance of the stratified treatment effect estimator. As the magnitude of α_1 increases, however, the covariance between X_2 and Y within strata and treatment groups increases and there is potential for improving the balance of X_2 within strata. Therefore, we expect the magnitude of the variance component V_3 to increase as α_1 increases.

Four values of α_1 are tried: -0.25, 0, 0.15 and 0.25. All other parameters are set to their default values. The changes in variance components as α_1 varies can be seen in Table 5.3 . As expected, the magnitude of both V_1 and V_3 increase with the magnitude of α_1.

Table 5.3: *Change in variance components as α_1 varies.*

α_1	-0.25	0	0.15	0.25
$n\,V_1$	702.66	626.82	637.82	662.02
$n\,V_2$	0.0007	0.0001	0.000008	0.00006
$n\,V_3$	-32.77	-1.84	-36.82	-77.90
$n\,V_4$	1.78	1.84	1.69	1.11

As before, V_1 accounts for most of the variance and V_2 is negligible. As the magnitude of α_1 increases, the imbalance of X_2 — the covariate related to outcome — increases, creating more potential for the estimated propensity score to reduce this within-stratum imbalance between treatment groups. We noted that when $\alpha_1 = 0$, we would not expect the estimation of the propensity score to affect the variance of the stratified treatment effect estimator. Table 5.3 shows that although, when $\alpha_1 = 0$, V_3 is non-zero and so we expect some benefit from estimating the propensity score, this is equal [1] to the additional variance incurred from the random error introduced by the estimation of the propensity score. Therefore, estimating the propensity score has no effect overall.

[1] In fact, the magnitude of the variance component V_4 is fractionally smaller than the magnitude of V_3.

Varying the parameter α_2

Since the outcome depends only on the covariate X_2 and treatment, the smallest outcome variance within strata and treatment groups would be achieved by stratifying on X_2 only. When α_2 is large in comparison with α_1, stratifying by the propensity score is almost equivalent to stratifying by X_2 only. As the magnitude of α_2 decreases, the propensity score is comparatively more influenced by the covariate X_1, increasing the within-stratum variance of X_2 which produces an increase in the outcome variance within stratum and treatment groups. Therefore, we expect the numerator of the variance component V_1 to decrease as the magnitude of α_2 increases.

We also need to consider the effect of changing α_2 on the population probabilities of being treated and in each stratum, the denominator of V_1. Large values of α_2 will greatly increase the range of the propensity score and produce very high or low probabilities of being treated and in some strata. This will increase the variance component V_1. This is likely to happen when α_2 is large. Therefore, we expect the variance component V_1 to initially decrease with the magnitude of α_2 but then to increase as α_2 gets large enough to produce strata containing extreme propensity scores.

Since the balance of the covariate X_2 within population strata should increase with the magnitude of α_2, we expect the magnitude of the variance component V_3 to decrease as α_2 increases.

Three values of α_2 are tried: 0.0075, 0.01 and 0.05. The resulting variance components are given in Table 5.4. These are in agreement with the discussion above.

Table 5.4: *Change in variance components as α_2 varies.*

α_2	0.0075	0.01	0.05
$n\,V_1$	724.20	637.82	701.34
$n\,V_2$	0.00004	0.000008	0.0033
$n\,V_3$	−65.00	−36.82	−2.53
$n\,V_4$	1.49	1.69	1.43

As before, Table 5.4 shows that V_1 accounts for most of the variance, V_2 is negligible, the magnitude of V_3 increases with the within strata variance of X_2, and V_4 is small but much larger than V_2.

5.2 Investigation of the convergence rates of the variance formulæ

Recall that we have shown that the variance of the stratified treatment effect estimator is given by

$$\mathbb{V}_k[\,\hat{\beta}^s\,] = V_1 + V_2,$$
$$\mathbb{V}_e[\,\hat{\beta}^s\,] = V_1 + V_2 + V_3 + V_4,$$

where the 'k' and 'e' refer to the propensity score being known and estimated, respectively (Theorems 3.1 and 3.2). Having calculated the value of these four variance components, V_1, V_2, V_3 and V_4, for a hypothetical example, and shown that each component behaves as expected, we now compare $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$ to empirical estimators of the same two variances. The latter should give us unbiased estimators of the finite sample variance of the stratified treatment effect estimator. The calculated variances are asymptotic results which should be close to the true variance for 'large enough' sample sizes. Thus comparing the empirical estimators with the calculated variances for various sample sizes should give some indication of how large a sample size is needed for the calculated variance formulæ to be approximately correct.

We calculate the 'true' values of $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$ as described in Section 5.1.1. Since we know the distribution of the data we can calculate the two variances mathematically, without having to estimate them from a dataset. We then obtain empirical estimates of the variance of the stratified treatment effect estimator, both when the propensity score is known and when it is estimated using a correctly specified logistic regression model, as follows. We simulate a dataset of size n and use it to obtain a stratified treatment effect estimate, stratifying on the 'true' propensity score. We then repeat the process 2,999 times, resulting in 3,000 stratified treatment effect estimates. These estimates can be used to approximate the sampling distribution of the stratified treatment effect estimator and so the variance of these estimates is an empirical estimate of $\mathbb{V}_k[\,\hat{\beta}^s\,]$. We obtain an empirical estimate of $\mathbb{V}_e[\,\hat{\beta}^s\,]$ in the same way, replacing the true propensity score by the estimated propensity score — estimated for each simulated dataset using a logistic regression model of treatment status on the covariates X_1 and X_2. These simulations are performed with the statistical software *Stata* [99] using the program given in Appendix D.1.

The four example situations we now consider are all variations on the example used previously (Section 5.1.1). In situations (a) and (b), all the conditions listed in Lemmas 3.1 and 3.2 and Theorems 3.1 and 3.2, that guarantee consistency and asymptotic normality, are satisfied. In examples (c) and (d), although all these conditions are still satisfied, we pick examples which are clinically plausible where one of the conditions mentioned is almost violated. This is to see both how easily the necessary conditions can be violated and what effect this has on the sample size necessary for the empirical estimates of variance to be approximately equal to the calculated variances.

In example situations (a) and (d) five strata are used. This is the typical number of strata used in applications. In the other two situations, (b) and (c), only two strata are used. In all examples, each strata contains an equal fraction of the sample, although the number of treated and untreated subjects in each stratum varies according to the distribution of the propensity score. All the empirical estimates of variances given in Figures 5.1 – 5.5 are based on 3,000 simulated datasets.

5.2.1 *Simulated example (a)*

In this example five strata are used. The propensity score is defined so that approximately 20% of each sample are treated. Details are as follows:

$$\begin{aligned}
\text{Outcome:} \quad & Y = 35 - 0.35\,X_2 + 2\,Z + \epsilon, \quad \epsilon \sim N(0, 10^2). \\
\text{Propensity score:} \quad & \ln\left(\frac{p_o(\mathbf{X})}{1 - p_o(\mathbf{X})}\right) = -2 + 0.15\,X_1 + 0.01\,X_2. \\
\text{Covariates:} \quad & \mathbb{P}(X_1 = 0) = 0.6, \\
& X_2 \,|\, X_1 = 0 \sim N(70, 10^2), \quad X_2 \,|\, X_1 = 1 \sim N(60, 15^2).
\end{aligned}$$

For each subject, receiving treatment ($Z = 1$) increases the outcome by 2. Therefore, in this example the population average causal treatment effect, β_o, is 2. The population stratified treatment effect, however, is defined as

$$\beta_o^s = \sum_{s=1}^{5} \{\mathbb{E}[Y \,|\, Z = 1,\, S_{so} = 1] - \mathbb{E}[Y \,|\, Z = 0,\, S_{so} = 1]\}/5 \;=\; 1.94,$$

where $S_{so} = 1_{[q_{(s-1)o} \le p_o(\mathbf{X}) < q_{so}]}$ is an indicator for the s^{th} population stratum. There is little residual confounding here and the population stratified treatment effect is

similar to the population average causal treatment effect. Since $\hat{\beta}^s$ is consistent for β_o^s, the estimate $\hat{\beta}^s$ should be a fairly good estimate of our estimand of interest, β_o.

Figure 5.1: *Theoretical and empirical variances of $\hat{\beta}^s$, for example (a), with the probability density function of the propensity score by treatment group.*

	$n\mathbb{V}_k[\hat{\beta}^s]$	$n\mathbb{V}_e[\hat{\beta}^s]$
Theoretical	637.8	602.7
Sample size (n)		
500	667.7	637.1
1,000	652.3	623.3
2,000	635.4	601.9

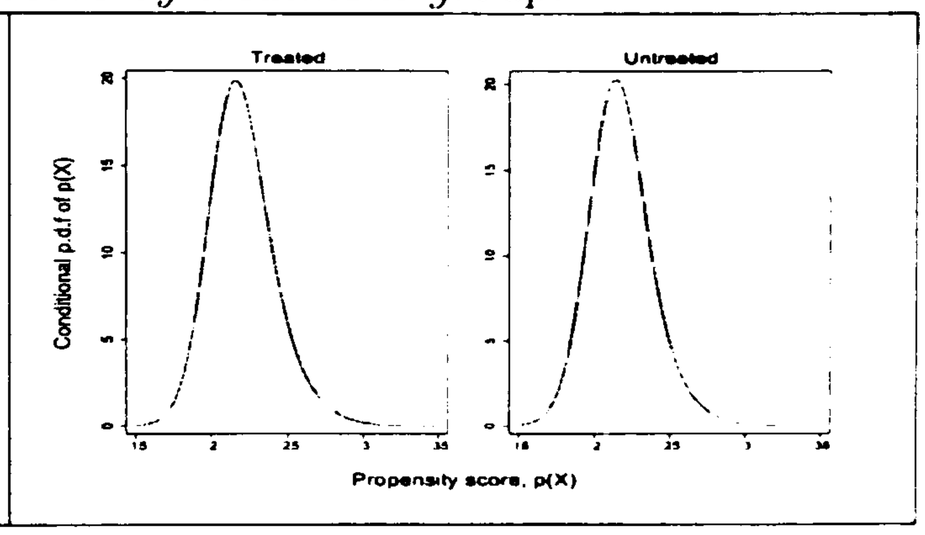


Figure 5.1 shows the probability density function of the propensity score by treatment group, calculated using the formula given earlier (5.2). Empirical estimates of both variances are given for three sample sizes: $n = 500$, $n = 1,000$ and $n = 2,000$. With the two smaller sample sizes, the empirical variance estimates using both the known and estimated propensity scores are greater than the calculated variances, as we would expect. Even at these smaller sample sizes, however, the difference between the two empirical variances is in the direction we expect — estimation of the propensity score reduces the variance — and approximately the correct magnitude. By $n = 2,000$ the empirical and theoretical results agree closely.

5.2.2 Simulated example (b)

In this example, two strata are used. The propensity score is defined so that subjects in both treatment groups have a wide range of propensity score values but the probability density function of the propensity score is not too small at the strata boundaries (see Figure 5.2). Details of the example are as follows:

$$\begin{aligned}
\text{Outcome:} \quad & Y = 35 + 0.3\,X_2 + 2\,Z + \epsilon, \quad \epsilon \sim N(0, 10^2). \\
\text{Propensity score:} \quad & \ln\left(\frac{p_o(\mathbf{X})}{1 - p_o(\mathbf{X})}\right) = -3 - 0.5\,X_1 + 0.05\,X_2. \\
\text{Covariates:} \quad & \mathbb{P}(X_1 = 0) = 0.6, \\
& X_2 \,|\, X_1 = 0 \sim N(70, 10^2), \quad X_2 \,|\, X_1 = 1 \sim N(60, 15^2).
\end{aligned}$$

As in the previous example, for each subject, receiving treatment ($Z = 1$) increases the outcome by 2. Therefore, the population average causal treatment effect, β_o, is

2. The population stratified treatment effect, is defined as

$$\beta_o^s = \sum_{s=1}^{2} \{\mathbb{E}[Y \mid Z = 1,\, S_{so} = 1] - \mathbb{E}[Y \mid Z = 0,\, S_{so} = 1]\}/2 \;=\; 3.03.$$

where $S_{so} = 1_{[q_{(s-1)o} \le p_o(\mathbf{X}) < q_{so}]}$ is an indicator for the s^{th} population stratum. There is a lot of residual confounding here — indicated by the difference between β_o and β_o^s. In practice more than two strata would be used for this example which would decrease the difference between β_o^s and β_o.

Figure 5.2: *Theoretical and empirical variances of $\hat{\beta}^s$, for example (b). with the probability density function of the propensity score by treatment group.*

	$n\mathbb{V}_k[\hat{\beta}^s]$	$n\mathbb{V}_e[\hat{\beta}^s]$
Theoretical	463.3	456.8
Sample size (n)		
500	478.4	477.2
1,000	482.9	479.3
2,000	461.1	457.5

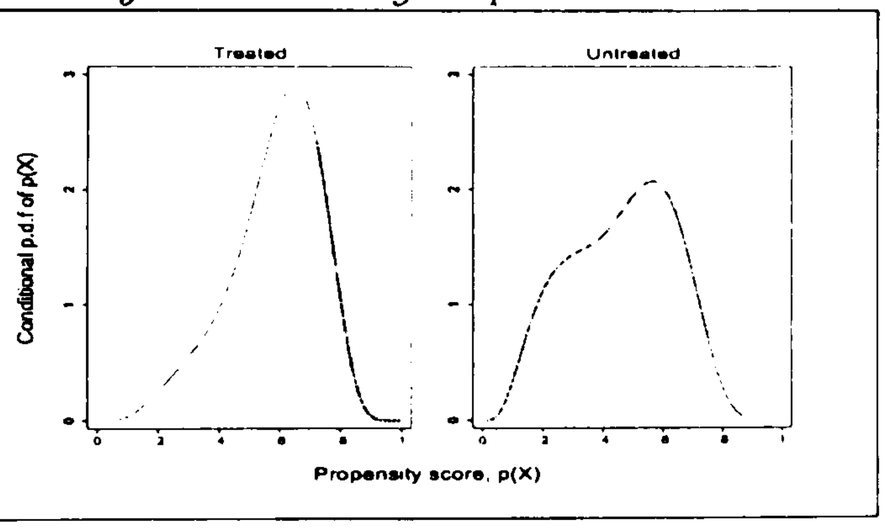


Figure 5.2 shows that for all sample sizes, the empirical variances are all larger than the theoretical variances by an amount that decreases with the sample size, as we would expect. As in the previous example, with a sample size of 2,000 the empirical results agree closely with the calculated variances.

5.2.3 Simulated example (c)

In this example, two strata are used. The propensity score is centred around 0.5 with a very small variance (see Figure 5.3). This example approximates a typical clinical trial situation where all subjects have approximately a 50% chance of receiving treatment. Details are as follows:

$$\begin{aligned}
\text{Outcome:} \quad & Y = 35 - 0.35\,X_2 + 2\,Z + \epsilon, \quad \epsilon \sim N(0, 2^2). \\
\text{Propensity score:} \quad & \ln\left(\frac{p_o(\mathbf{X})}{1 - p_o(\mathbf{X})}\right) = -0.1 + 0.01\,X_1 + 0.005\,X_2. \\
\text{Covariates:} \quad & \mathbb{P}(X_1 = 0) = 0.6, \\
& X_2 \mid X_1 = 0 \sim N(70, 10^2), \quad X_2 \mid X_1 = 1 \sim N(60, 15^2).
\end{aligned}$$

Again, for each subject, receiving treatment ($Z = 1$) increases the outcome by 2 and so the population average causal treatment effect, β_o, is 2. The population stratified treatment effect is defined as

$$\beta_o^s = \sum_{s=1}^{2} \{\mathbb{E}[Y \mid Z = 1, S_{so} = 1] - \mathbb{E}[Y \mid Z = 0, S_{so} = 1]\}/2 \;=\; 1.88,$$

where $S_{so} = 1_{[q_{(s-1)o} \leq p_o(\mathbf{X}) < q_{so}]}$ is an indicator for the s^{th} population stratum. There is little residual confounding in this example so we expect our estimator, $\hat{\beta}^s$, to be a fair estimate of the parameter of interest, β_o.

As we mentioned previously, this example was chosen since it nearly violates one of the conditions necessary for asymptotic normality of the stratified treatment effect estimator, when the propensity score is estimated. In particular, condition *(iii)* of Lemma 3.2 states that the derivative of the probability density function of the propensity score with respect to each propensity score parameter must be bounded. Table 5.5 shows the maximum of the derivative of the probability density function of the propensity score with respect to α_2. This shows that in example (d), the maximum derivative is finite and hence satisfies the condition. It is, however, much larger than the maximum derivative in any of the other hypothetical examples chosen. Therefore, although the variance $n\,\mathbb{V}_e[\hat{\beta}^s]$ is asymptotically valid in this example, we expect the empirical variance to converge more slowly to $n\,\mathbb{V}_e[\hat{\beta}^s]$ than it did for the previous two examples. Note that we do not expect any problems with the convergence of $n\,\mathbb{V}_k[\hat{\beta}^s]$ since the condition that is almost violated is only necessary when the propensity score is estimated from the data.

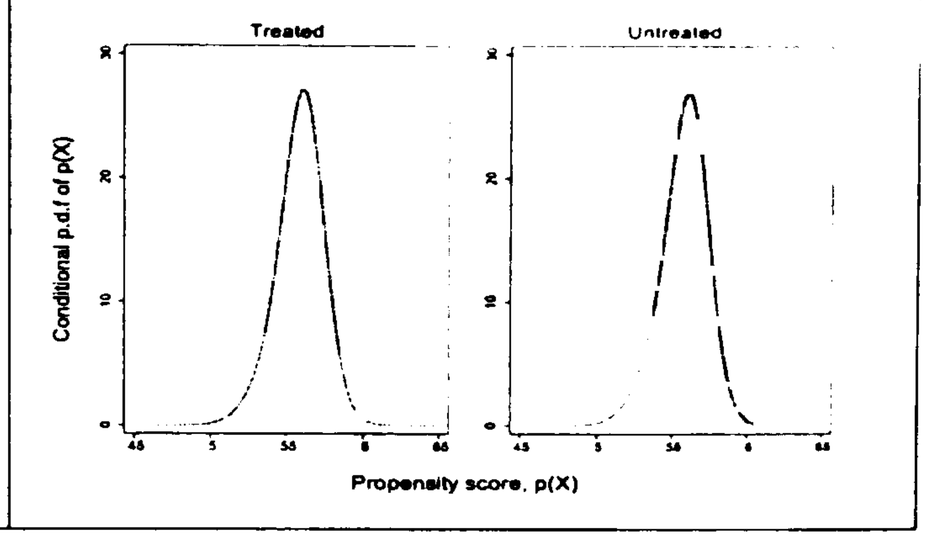
Table 5.5: *Maximum value of the derivative of the probability density function of the propensity score with respect to* α_2.

Simulated example	α_2	$\text{Maximum}_{p\in(0,1)} \left\{ \frac{\partial f_p(p)}{\partial \alpha_2} \right\}$
(a)	0.01	6,312.5
(b)	0.05	248.1
(c)	0.005	21,369.9
(d)	0.085	3.1

In this example, the calculated variances suggest that estimation of the propensity score should not affect the variance. However, in practice, Figure 5.3 shows that estimation of the propensity score results in a decrease in variance for all sample sizes up to $n = 50,000$. The decrease in variance associated with estimating the propensity

Figure 5.3: *Theoretical and empirical variances of $\hat{\beta}^s$, for example (c), with the probability density function of the propensity score by treatment group.*

	$n\mathbb{V}_k[\hat{\beta}^s]$	$n\mathbb{V}_e[\hat{\beta}^s]$
Theoretical	50.2	49.9
Sample size (n)		
500	50.3	40.9
5,000	49.5	45.4
10,000	49.3	46.1
50,000	50.7	50.6

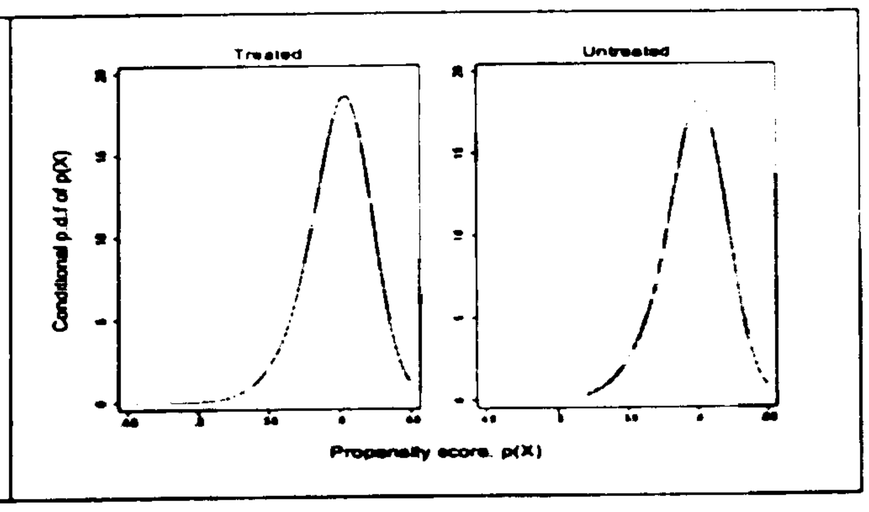


score is reduced as the sample size increases. As predicted, the empirical variance is very close to the calculated variance when the propensity score is known, even for very small sample sizes. When the propensity score is estimated, however, very large sample sizes are needed for the empirical variance to be close to the calculated variance. This shows that almost violating the condition discussed above has drastically slowed the convergence rate.

In order to investigate the effect of this large derivative on the empirical variances, the simulation study was repeated for this example situation, replacing $\alpha_2 = 0.005$ by $\alpha_2 = 0.0075$. This produces only a small change in the propensity score distribution as is shown in Figure 5.4. The maximum derivative of the probability density function with respect to α_2 then becomes 9,876.3 — still larger than all the other examples but smaller than it was. This small change in the propensity score parameters considerably improves the speed of convergence. Figure 5.4 shows that with a sample size of $n = 2,000$, the estimation of the propensity score still appears to reduce the variance despite the calculated variances predicting no reduction in variance but that with a sample size of $n = 5,000$, the theoretical and empirical variances agree closely.

Figure 5.4: *Theoretical and empirical variances of $\hat{\beta}^s$, for example (c) with $\alpha_2 = 0.0075$, with the probability density function of the propensity score by treatment group.*

	$n\mathbb{V}_k[\hat{\beta}^s]$	$n\mathbb{V}_e[\hat{\beta}^s]$
Theoretical	51.4	51.3
Sample size (n)		
2,000	52.7	46.8
5,000	51.6	49.9



5.2.4 Simulated example (d)

In this example five strata are used. The propensity score has a bimodal distribution as can be seen in Figure 5.5. Details of the example are as follows:

$$\text{Outcome:} \qquad Y = 8 - 4\,X_2 + 2\,Z + \epsilon, \quad \epsilon \sim N(0, 8^2).$$

$$\text{Propensity score:} \qquad \ln\left(\frac{p}{1-p}\right) = -1 + 1.5\,X_1 + 0.085\,X_2.$$

$$\text{Covariates:} \qquad \mathbb{P}(X_1 = 0) = 0.6,$$

$$X_2 \,|\, X_1 = 0 \sim N(5, 3^2), \quad X_2 \,|\, X_1 = 1 \sim N(10, 4^2).$$

As in the previous example, for each subject, receiving treatment ($Z = 1$) increases the outcome by 2. Therefore, the population average causal treatment effect, β_o, is 2. The population stratified treatment effect, is defined as

$$\beta_o^s = \sum_{s=1}^{5} \{\mathbb{E}[Y \,|\, Z = 1,\, S_{so} = 1] - \mathbb{E}[Y \,|\, Z = 0,\, S_{so} = 1]\}/5 = 0.88,$$

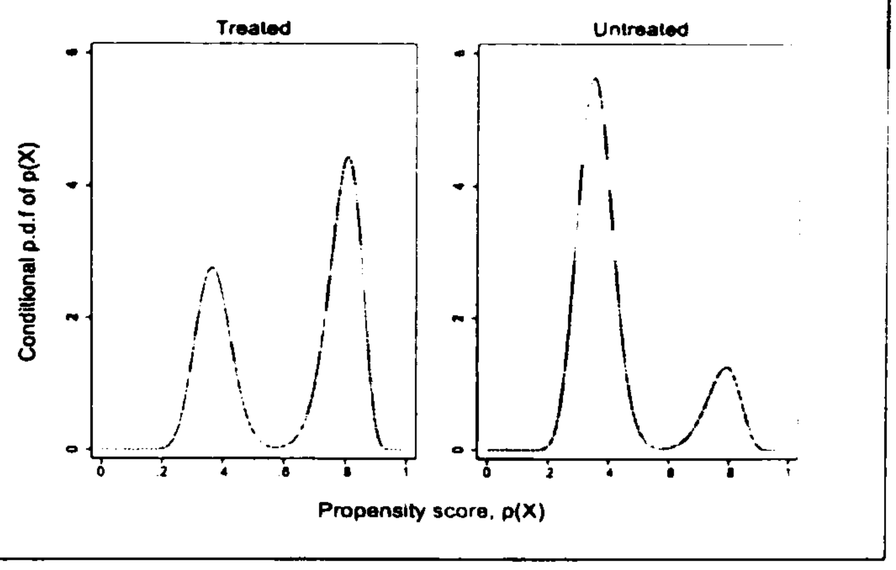
where $S_{so} = 1_{[q_{(s-1)o} \le p_o(\mathbf{X}) < q_{so}]}$ is an indicator for the s^{th} population stratum. There is a lot of residual confounding in this example — β_o is far from β_o^s. In practice, since the distribution of the propensity score is so different in the two treatment groups, more than five strata would be used to analyse these data if the sample size allowed.

This example was also chosen to almost violate one of the conditions required for the validity of the calculated variances. In particular, condition *(ii)* of Theorem 3.1 states that the probability density function of the propensity score must be non-zero at each population strata boundary. This condition is necessary for the validity of both $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$. In this hypothetical situation, the probability density function of the propensity score at the third population strata boundary is very low, $f_p(q_{3o}) = 0.028$. We therefore expect that the empirical estimates of both variances will converge slowly to the calculated variances.

The calculated variances suggest that estimation of the propensity score will decrease the variance by about 40. Figure 5.5 shows that in practice, estimation does decrease the variance, although this decrease is less than expected. However, all of the empir-

Figure 5.5: *Theoretical and empirical variances of $\hat{\beta}^s$, for example (d), with the probability density function of the propensity score by treatment group.*

	$n\mathbb{V}_k[\hat{\beta}^s]$	$n\mathbb{V}_e[\hat{\beta}^s]$
Theoretical	630.9	590.8
Sample size		
500	729.4	701.4
5,000	711.6	689.6
10,000	721.9	703.1
50,000	752.2	731.5



ical estimates are much larger than the calculated variances, even with a sample size of $n = 50,000$. So as expected, the convergence rate of both variances is very slow.

Figure 5.6 shows histograms of the four estimated strata boundaries from 3,000 simulated datasets, each containing 5004 [2] subjects. The probability density function of the propensity score at each population strata boundary is shown above the histograms. The histograms show that although the empirical sampling distributions of three of the strata boundaries appear to be approximately normal, the empirical sampling distribution of the third strata boundary is bi-modal. The probability density function of the propensity score is very low at this strata boundary, $f_p(q_{3o}) = 0.028$, which accounts for this non-normality.

If the population strata boundary falls in an area with low probability density function, whilst this may make it hard to estimate the strata boundary, we might expect the chance of misclassification of propensity scores to be quite low. Thus we might expect this to have relatively little impact on the estimator or its variance. However, as predicted in discussions in Chapter 3 and confirmed in Figure 5.5, this does not appear to be the case. We now investigate further the reasons why we might expect the variance of the stratified treatment effect estimator to be affected by one of the population strata boundaries falling on a point where the propensity score probability density function is low. We concentrate on the simpler case where the propensity score is a known function of the observed covariates.

[2] We use 5004 subjects now and 5003 a little later because the emphasis here is on the estimated strata boundaries, and so choosing a sample size such that all strata boundaries fall exactly on an observation prevents unnecessary interpolation.

Figure 5.6: *Histograms of the 4 estimated strata boundaries from 3,000 simulated datasets from example (d) with sample size n=5,004, with the probability density function of the propensity score at each population strata boundary.*

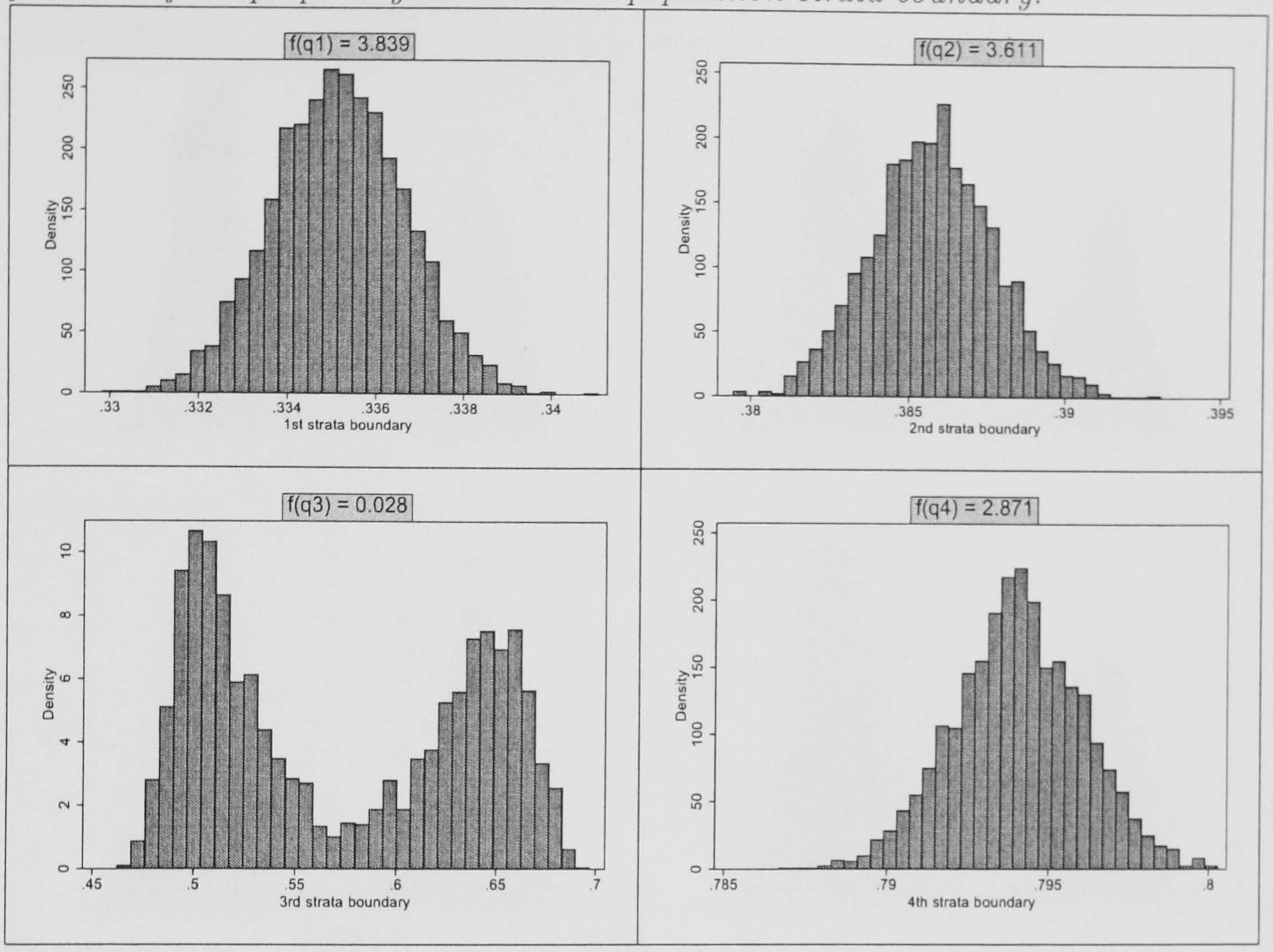
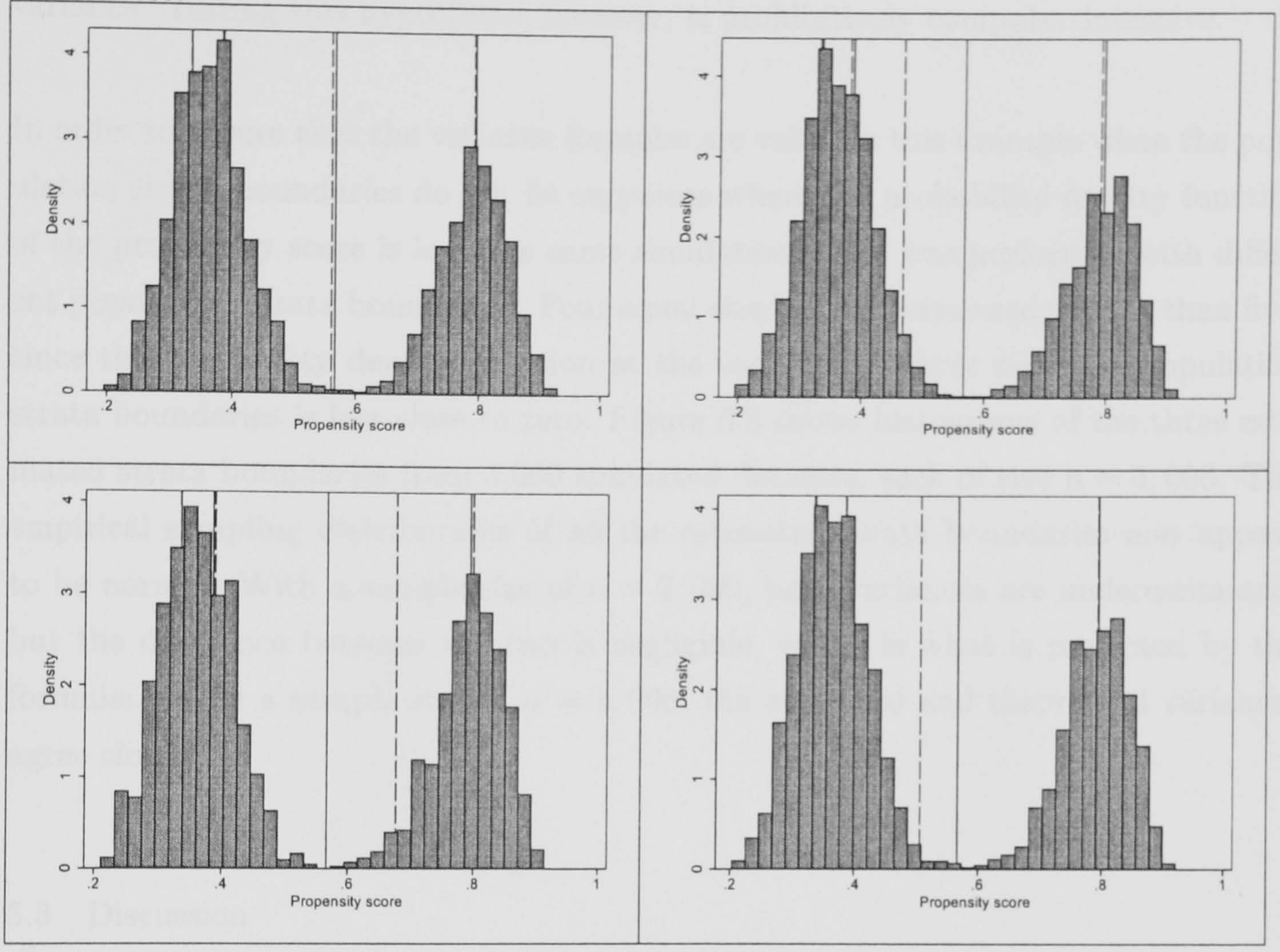


Figure 5.7 shows histograms of the (true) propensity scores from four simulated datasets from example (d) each containing 2,000 subjects. Four solid vertical lines in each histogram show the population strata boundaries. The four dashed lines in each histogram show the estimated strata boundaries. For each of the four strata boundaries only around 40 subjects are misclassified. At the first, second and fourth population strata boundaries, the probability density function of the propensity score is quite high so the 40 or so subjects either side of the population strata boundary have a propensity score similar to the population strata boundary. Therefore, the disparity between the estimated and true strata boundaries, other than the third, is small. In contrast, the 40 subjects either side of the third population strata boundary have a much greater range of propensity scores due to the low probability density function at this point, which can lead to a much greater disparity between the estimated and true third strata boundary.

Figure 5.7 shows that by estimating a strata boundary in an area with low probability density function we can greatly increase or decrease the range of the propensity score

Figure 5.7: *Histograms of four simulated datasets from example (d) with sample size n=2,000. The four solid vertical lines in each histogram show where the population strata boundaries lie. The four dashed vertical lines represent the estimated strata boundaries.*



within the strata defined by that strata boundary. Since the propensity score is the probability of being treated, unless the population strata boundary falls on 0.5 and the disparity between that and the estimated strata boundary is quite small, we can expect the numbers of treated and untreated subjects that are misclassified to be different. If also, as is usual, the outcome is correlated with the propensity score, we can expect this differential misclassification to change the within-stratum estimates of treatment effect, and hence the stratified treatment effect estimate.

Even in larger samples, this misclassification will occur. However, in large enough samples, the error due to misclassification will be well estimated by the second variance component. In smaller samples, the empirical distribution of the propensity score may be very dissimilar to the true probability density function of the propensity score in areas of low density. This will lead to the misclassification error being even higher than expected. We therefore might predict that when a strata boundary falls on an area of low density, as in example (d), the finite-sample variance would

be higher than that predicted by the asymptotic variance formulæ. This, as we have seen, is exactly the case. In infinite or 'large enough' samples it is supposed that the formulæ $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$, would give be very similar to empirical estimates of variance. Testing this hypothesis, however, is prohibitively computer-intensive.

In order to ensure that the variance formulæ are valid for this example when the population strata boundaries do not lie on points where the probability density function of the propensity score is low, the same simulation study was performed with different population strata boundaries. Four equal sized strata were used, rather than five, since the probability density function at the each of the three resulting population strata boundaries is less close to zero. Figure 5.8 shows histograms of the three estimated strata boundaries from 3,000 simulated datasets, each of size $n = 5,003$. The empirical sampling distributions of all the estimated strata boundaries now appear to be normal. With a sample size of $n = 2,000$, both variances are underestimated, but the difference between the two is negligible, which is what is predicted by the formulæ. With a sample size of $n = 5,000$, the empirical and theoretical variances agree closely.

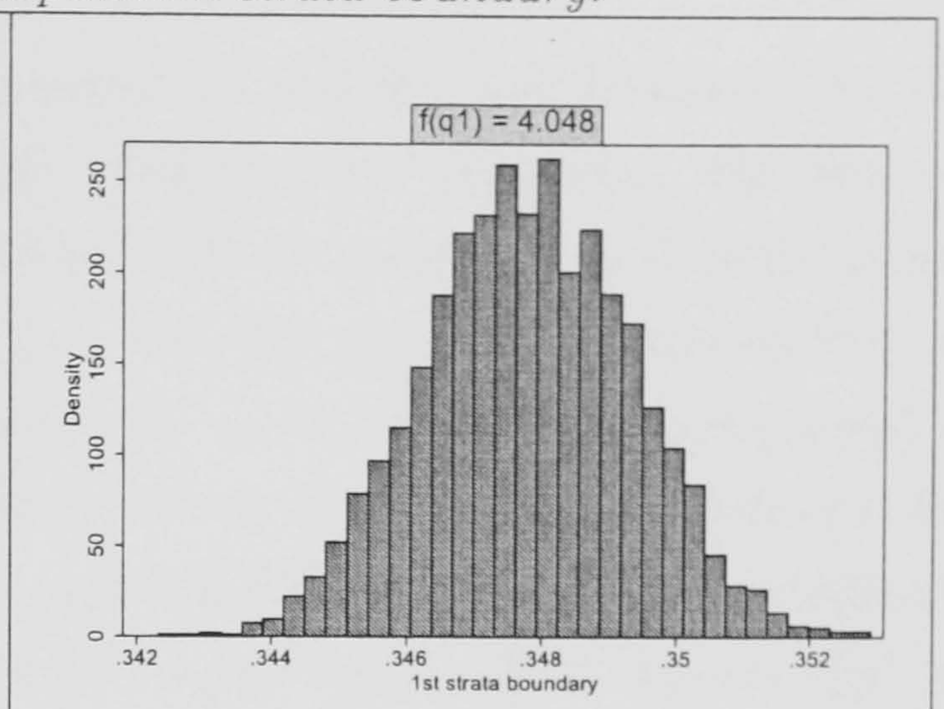
5.3 Discussion

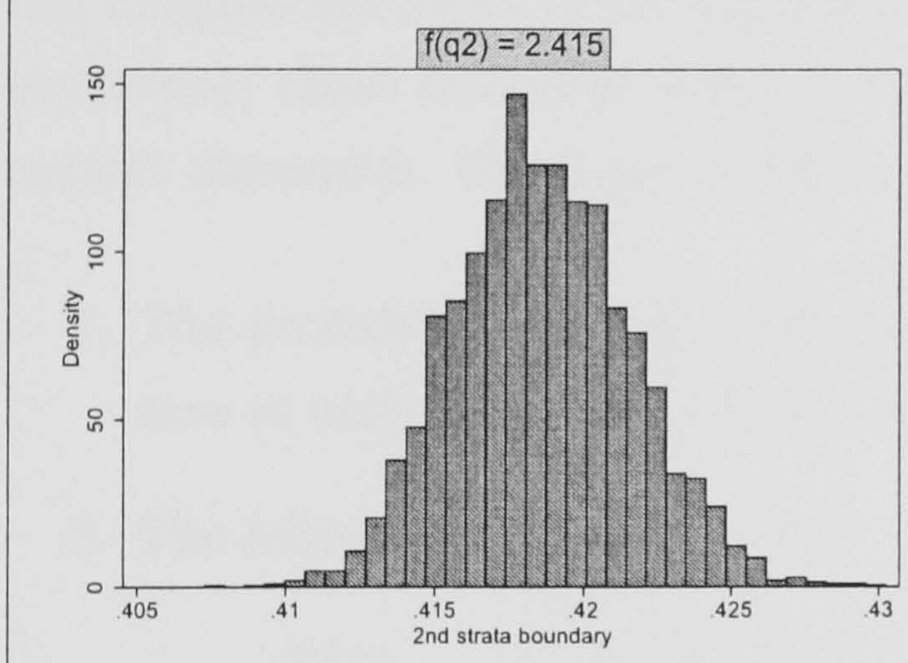
In Chapter 3, we calculated the variance of the stratified treatment effect estimator, assuming that the propensity score is: (i) a known function of the observed covariates, and (ii) estimated using a correctly specified logistic regression model. We denoted these variances by $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$, respectively, and expressed them in terms of four variance components, V_1, V_2, V_3 and V_4. We discussed the mathematical meaning of these four components (Section 3.4). In this chapter, we therefore began by calculating V_1, V_2, V_3 and V_4 for a simple hypothetical situation, varying the example parameters one at a time, finding that the resulting change in variance components agreed with our intuition, gained from the discussion mentioned above.

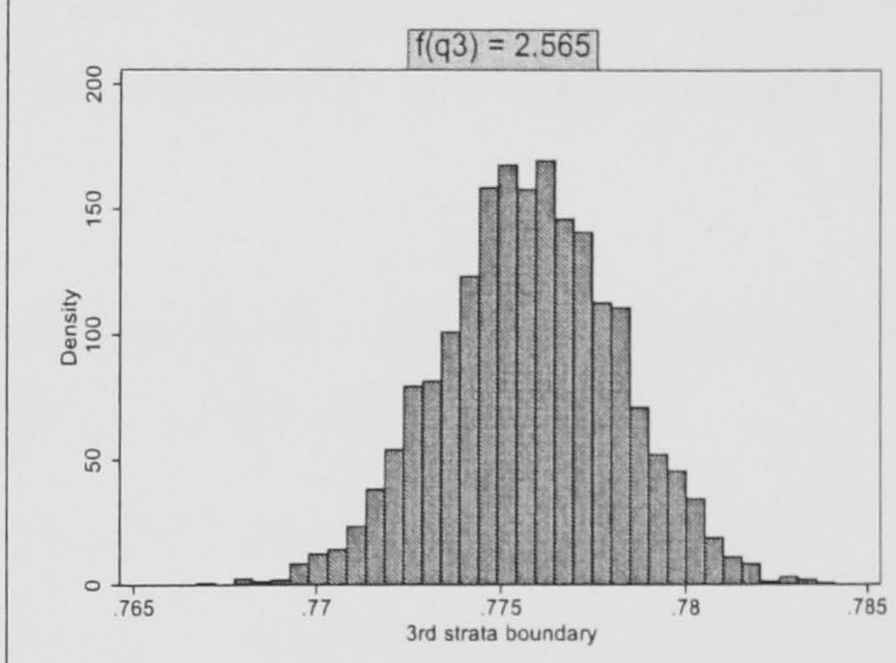
We then proceeded to investigate convergence rates for the two variance formulæ, $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$. Conditions under which these two variances are asymptotically valid are specified in Lemmas 3.1 and 3.2 and Theorems 3.1 and 3.2. We first looked at how large a sample size is needed to ensure that the two variance formulæ are valid when these conditions are all satisfied, and then considered the effect of almost

Figure 5.8: *Theoretical and empirical variances of $\hat{\beta}^s$, for example (d) using four strata, with histograms of the estimated strata boundaries, with the probability density function of the propensity score at each population strata boundary.*

	$n\mathbb{V}_k[\,\hat{\beta}^s\,]$	$n\mathbb{V}_e[\,\hat{\beta}^s\,]$
Theoretical	671.7	671.6
Sample size (n)		
2,000	633.8	632.5
5,000	674.8	674.4







violating one of the conditions. For four hypothetical situations, variations on the example used previously, we calculated the two variances — as above, these were calculated mathematically rather than estimated from a dataset — and compared these variances to empirical estimates of the same variances, obtained using various sample sizes. Our conclusions were as follows. For the first two examples, (a) and (b), all conditions necessary for the validity of the two variance formulæ were satisfied. We found that for sample sizes of $n = 2,000$, the empirical and calculated variances agreed well. Example (c) almost violated one of the conditions for the validity of $\mathbb{V}_e[\,\hat{\beta}^s\,]$ — that the derivative of the probability density function of the propensity score with respect to each propensity score parameter should be bounded — and as a result, although the calculated and empirical estimates of $\mathbb{V}_k[\,\hat{\beta}^s\,]$ agreed for all sample sizes considered, it took a sample size of $n = 50,000$ for the calculated and empirical estimates of $\mathbb{V}_e[\,\hat{\beta}^s\,]$ to agree. Similarly, example (d) was chosen to almost violate a condition required for the validity of both variances — the probability density function of the propensity score was very close to zero at one of the population strata

boundaries. In this case, all the empirical and calculated variances were still dissimilar with sample sizes of $n = 50,000$.

Looking at the conditions specified in Lemmas 3.1 and 3.2 and Theorems 3.1 and 3.2, we can see that the conditions broadly fall into two categories. The first set merely ensure that the problem is well-defined, and the second set deal with the rate of convergence of the asymptotic variances. The first set cover conditions such as the population probability of being treated and in the s^{th} stratum being equal to neither zero nor r_s, for $s = 1, ..., K$, without which the population quantity being estimated, β_o^s, would be undefined. It also includes conditions such as the continuity of the probability density function of the propensity score, which ensures that we can measure the effect of the estimation of the propensity score. The second set of conditions, those that deal with the rate of convergence, are more relevant for our current discussion. These are as follows.

1. The probability density function of the propensity score, $f_p(\cdot)$, must be non-zero at each population strata boundary, q_{so}, for $s = 1, ..., K-1$.

2. The following functions must be bounded for all $p \in (0,1)$ and $t = 0, 1$:

$$f_p(p), \qquad \mathbb{E}[Y \mid Z = t,\, p_o(\mathbf{X}) = p], \qquad \mathbb{E}[Y^2 \mid Z = t,\, p_o(\mathbf{X}) = p].$$

3. The following functions must be bounded for $p \in (0,1)$, for $t = 0, 1$ and $k = 1, ..., m$:

$$\frac{\partial f_p(p)}{\partial \alpha_k}, \quad \frac{\partial}{\partial \alpha_k}\left\{\mathbb{E}[Y \mid Z = t,\, p(\mathbf{X}) = p]\right\}, \quad \frac{\partial}{\partial \alpha_k}\left\{\mathbb{E}[Y^2 \mid Z = t,\, p(\mathbf{X}) = p]\right\}.$$

In particular, we are concerned with the effect of almost, rather than absolutely, violating these conditions. We have seen that almost violating condition 1 means that a larger sample size is necessary for the validity of both variances. Almost violating condition 2 would similarly affect the convergence rate of both variances, whereas almost violating condition 3 only affects the convergence rates of the variance when the propensity score is estimated.

In practice, we would not expect condition 2 to be a problem since the conditional expectations would have to be extremely large to pose a problem. Condition 1 could be detected by graphing the estimated density of the propensity score, and could be

solved, if necessary, by redefining the strata, as we did in this chapter. Condition 3 is harder both to detect and solve. We can expect this problem to occur when a small change in a propensity score parameter greatly changes the shape of the probability density function of the propensity score. This is what we found in example (c). It is not clear, at present, whether it is possible to detect this problem in a dataset.

So far, we have only considered hypothetical examples where the distribution of the data is known and so we have calculated the variances $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$ mathematically. However, in practice, we will not be able to do this and must estimate these variances from a sample dataset. The next chapter, therefore, considers methods of estimating the four variance components, V_1, V_2, V_3 and V_4, and hence the two variances, from a sample dataset.

Chapter 6

Estimating the variance of the stratified treatment effect estimator

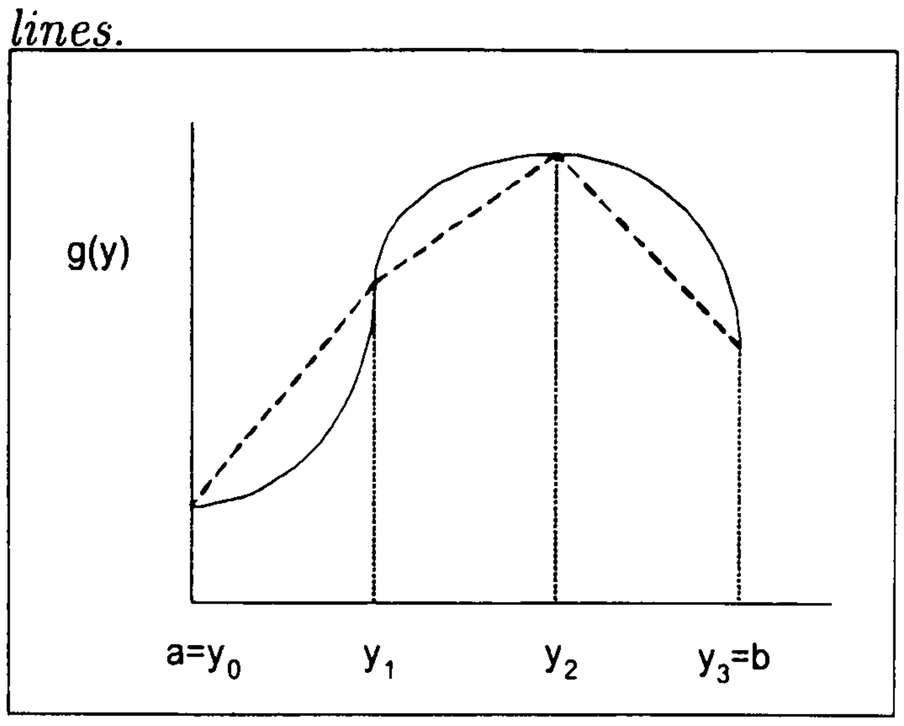
We have calculated formulæ for the variance of the stratified treatment effect estimator when the propensity score is: (i) a known function of the observed covariates, and (ii) estimated using a correctly specified logistic regression model, denoted by $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$, respectively. We expressed these two variances in terms of four variance components, V_1, V_2, V_3 and V_4. In practice, we will not be able to calculate these variance components mathematically as we did in Chapter 5. We now, therefore, consider the issue of estimating the variance of the stratified treatment effect estimator from a sample dataset. We will see that whilst the variance components V_1 and V_3 are easily estimated, and V_2 can be expected to be negligible, the component V_4 is generally non-negligible and complex to estimate. To solve this estimation problem, we turn to the theory of kernel density estimation.

We begin this chapter by reviewing the mathematical tools that we use later in the chapter. We briefly describe the application of the trapezium rule to estimate a definite integral, after which we review the theory of kernel density estimation and kernel regression, applying the theory to our particular problem. We then demonstrate the estimation of the four variance components for two of the hypothetical datasets described in Chapter 5. We finish by using our variance estimators to construct confidence intervals for the stratified treatment effect estimator.

6.1 Some mathematical tools

We now provide an overview of the mathematical tools that will be used later in this chapter. We begin by showing how the trapezium rule can be used to estimate definite integrals and then move on to discuss the non-parametric methods of kernel density estimation and kernel regression.

Figure 6.1: *The trapezium rule. The area under the solid curve is estimated by the area under the dashed lines.*



6.1.1 Numerical integration using the trapezium rule

Suppose we wish to calculate the following definite integral of some function, $g(y)$,

$$I = \int_a^b g(y)\, dy.$$

When the integral is analytically intractable we cannot calculate I exactly but must use numerical techniques to obtain an approximation. One such technique is the trapezium rule [105]. We illustrate this method with a simple example before giving the general formula.

Figure 6.1 shows the function $g(y)$ from a to b. We have partitioned the interval $[a, b]$ into three sub-intervals, each of which has width $(b-a)/3$. Within the first of these sub-intervals we can estimate the integral of the function $g(y)$ by the area under the straight (dashed) line which connects the points $g(y_1)$ and $g(y_o)$. Since the area under this dashed line is a trapezium it is equal to

$$\frac{(b-a)}{3}\ \frac{(g(y_0)+g(y_1))}{2}.$$

Repeating this for each sub-interval, we find that the total area under the dashed lines is

$$\frac{(b-a)}{3}\ \frac{(g(y_0)+2g(y_1)+2g(y_2)+g(y_3))}{2}.$$

This can be used as an approximation of the integral I. In order to improve the approximation we could increase the number of sub-intervals. If we divide the interval

$[a, b]$ into M sub-intervals, defined by $a = y_0 < y_1 < ... < y_{M-1} < y_M = b$, each with width (b - a)/M, then we can estimate the integral I by

$$\hat{I} = \frac{(b-a)}{2M} \{g(y_0) + 2g(y_1) + ... + 2g(y_{M-1}) + g(y_M)\}.$$

6.1.2 Kernel density estimation

We now review another topic — the non-parametric method of kernel density estimation. Suppose we wish to estimate the probability density function of a continuous variable X, denoted by $f(X)$, from a sample of data, $\{X_i\}$ for $i = 1, ..., n$, drawn independently from the population. One way of doing this would be to assume a parametric form for X and to use the data to estimate the unknown parameters of the specified distribution. For example, we might assume that X is normally distributed with unknown mean and variance, which we would then estimate from the data. This approach assumes that we are able to correctly specify the parametric distribution of X, despite empirical evidence that continuous data do not always follow a classical parametric distribution (see examples in [96]). Conversely, non-parametric estimators of the probability density function avoid the necessity of specifying a particular distribution for X, which results in a more flexible, although less statistically powerful, approach. We consider the simplest non-parametric density estimator — the histogram — and show how this idea can be extended to produce a smooth estimator of the probability density function of a continuous variable from a sample dataset. Since, in our problem, we will need to estimate the derivatives of a probability density function, it is necessary to obtain a differentiable, and hence smooth, estimator of the density function.

The histogram

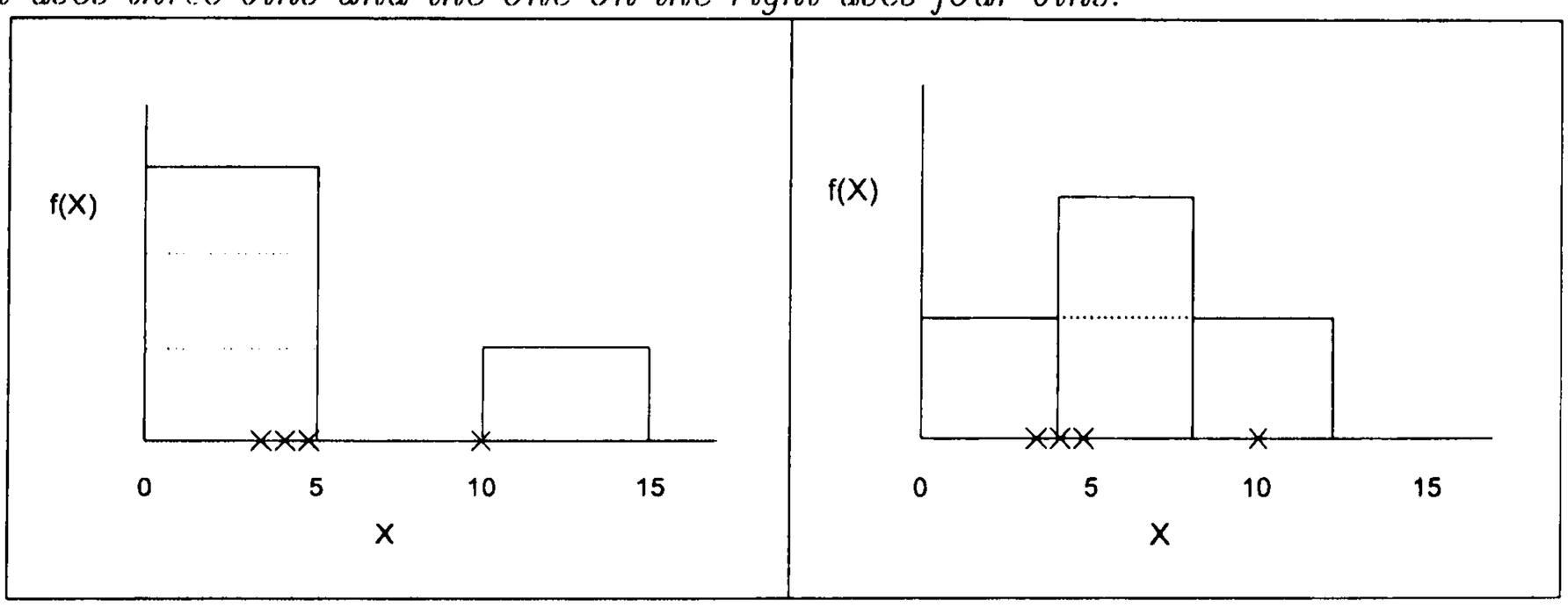
In order to construct a histogram for a sample of data, $\{X_i\}$ for $i = 1, ..., n$, we would proceed as follows. First we choose the number of bins — equal width sub-intervals that partition the interval in which the observed data lie — which we denote by N. Too few bins inadequately represent the characteristics of the dataset and too many place too much emphasis on random error in the data. We call this oversmoothing and undersmoothing the data, respectively. We then choose a starting point, x_{min}, which is less than or equal to the smallest sampled value of X, and a finishing point,

x_{max}, which is greater than or equal to the largest sampled value of X. Then the width of each bin is $h = (x_{max} - x_{min})/N$. If we let $B(X)$ represent the bin in which the point X lies, then the density estimator at X is

$$\hat{f}(X) = \frac{1}{nh} \sum_{i=1}^{n} 1_{[X_i \in B(X)]}.$$

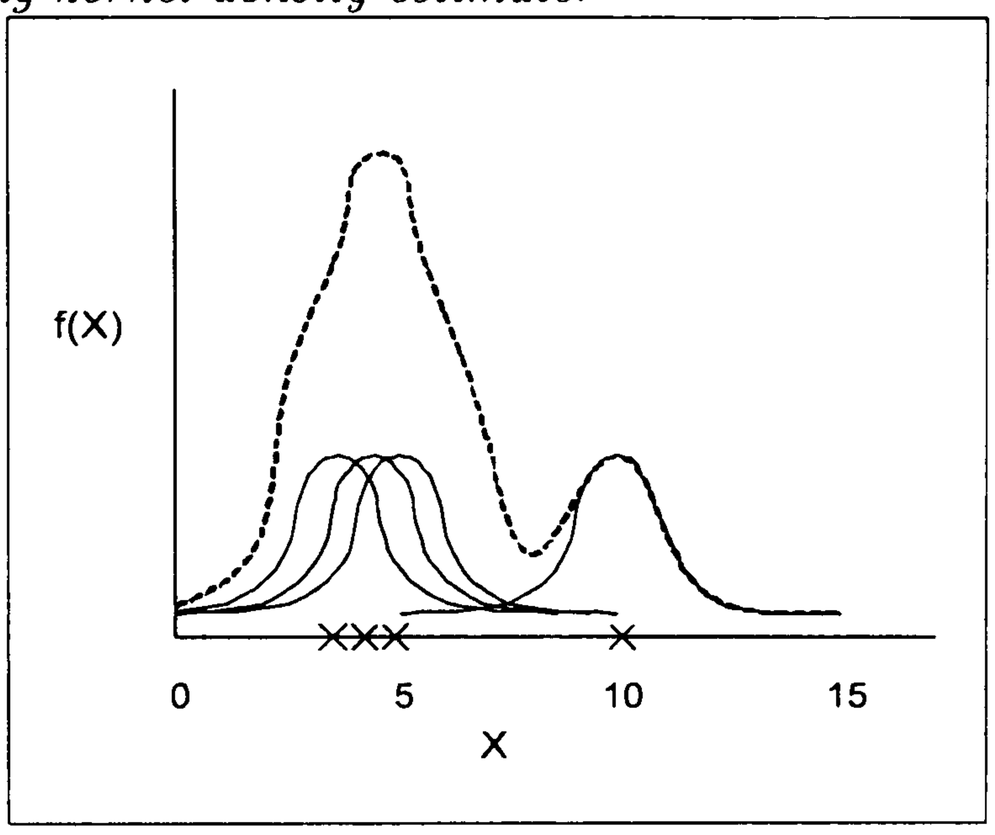
We can imagine the construction of the histogram as follows. Each observed data point, X_i, is given a box of width h and height $1/nh$. Note that this box has area $1/n$. In the histogram, X_i's box is placed in the bin to which X_i belongs. If a particular bin contains more than one observed value of X then the contributions of all these observed data points are added together. Or, in other words, the boxes are stacked on top of each other. The density estimator $\hat{f}(X)$ is then the height of the stacked boxes at point X.

Figure 6.2: *Two histograms of the dataset* $\{3.3, 4.1, 4.9, 10\}$. *The histogram on the left uses three bins and the one on the right uses four bins.*



We now consider a simple example. Suppose we collect a sample dataset containing four observations: $\{3.3, 4.1, 4.9, 10\}$. Figure 6.2 shows two possible histograms that we could use to represent this dataset. The first uses three bins of width five, starting from zero. The second uses four bins of width four, again starting from zero. The first histogram might suggest that X has a bimodal distribution, whereas the second suggests a symmetric distribution. These two different representations of the data highlight an important disadvantage of the histogram: its dependence on the choice of both the starting point and the bin-width. A second disadvantage, which is more relevant to our problem, is that the density estimator — the histogram — is not smooth and therefore not differentiable.

Figure 6.3: *A kernel density estimate using the dataset* $\{3.3, 4.1, 4.9, 10\}$. *The solid lines indicate the four normal kernels for the four observed values. The dashed line indicates the resulting kernel density estimate.*



A kernel density estimator

The kernel density estimator of a probability density function is an extension of the histogram idea above. Rather than allocating each observed data point a box of area $1/n$, which is that observation's contribution to the density estimator, each observed data point is given a 'kernel' of area $1/n$. This kernel can be a box, a triangle, a bell-shaped normal curve, or a number of other shapes. With the histogram, an observation's box is placed, rather abitrarily, in the bin to which it belongs and so the contribution to the density estimator is usually not symmetric about the observed value. By contrast, the kernel is centred about the observed data point. Then to get the density estimator at a particular point, we merely add up the heights of all the kernel shapes at that point.

To illustrate this idea, we revisit the simple example dataset $\{3.3, 4.1, 4.9, 10\}$. Using a bell-shaped normal distribution centred around the observed datapoints, with some pre-specified variance h^2 we obtain the density estimate shown in Figure 6.3.

A general kernel density estimator using a normal kernel, centred around the observed values, each with variance h^2, can be written as

$$\hat{f}(X) = \frac{1}{n} \sum_{i=1}^{n} \frac{e^{-\frac{(X-X_i)^2}{2h^2}}}{\sqrt{2\pi h^2}}. \tag{6.1}$$

The scaling factor $1/n$ ensures that each kernel has area $1/n$, and so each observed data point contributes the same amount to the density estimator. If we write the density of the standard normal distribution as $K(u) = e^{-u^2/2}/\sqrt{2\pi}$, then the probability density function estimator (6.1) can be written as

$$\hat{f}(X) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X - X_i}{h}\right). \tag{6.2}$$

This is often called the Parzen-Rosenblatt kernel density estimator [72, 87]. The function $K(\cdot)$ is called the kernel function. There are many other distributions that we could use in place of the normal distribution. However, from (6.2) we see that the density estimator $\hat{f}(X)$ will inherit properties of the kernel such as smoothness and differentiability. In our problem, these two properties will be very important, and most other common kernels do not share them. Since it has been shown that the choice of kernel makes little difference to the density estimate [63], we concentrate solely on the normal kernel function.

The square root of the variance of the normal densities used to construct the probability density estimator, h, is referred to as the bandwidth. The choice of this parameter determines the degree of smoothing applied to the data. As with histograms, it is important to smooth the data enough to remove most of the random error but not so much that genuine characteristics of the underlying data are missed. An h that is too large will oversmooth the data and an h that is too small will undersmooth the data. Intuitively, our choice of h should depend on two things. Firstly, if our data had very large variance we would wish to smooth it more than if there was little variance. Therefore, we expect h to increase with the variance of X, which we will call σ_x^2. Secondly, as our sample size grows we wish to give less weight to observations further from the point of interest and so we want h to decrease as n increases. One way of choosing h is to minimise the mean integrated squared error (MISE) of the kernel density estimator, which is

$$MISE = \int_X \mathbb{E}\left[\hat{f}(x) - f(x)\right]^2 dx + \int_X \mathbb{V}\left[\hat{f}(x)\right] dx.$$

When the density $f(X)$ is normal, this can be shown to be minimised asymptotically by the choice $h \approx 1.06\, n^{-1/5}\, \sigma_x$ [96]. As desired, this choice of h increases with the variance of the data and decreases with the sample size. Estimating σ_x^2 by a sample variance and plugging the resulting bandwidth into (6.2) gives the required density estimator.

Kernel regression

Having seen how we might obtain a non-parametric estimator of the probability density function of a continuous variable, we now proceed to a slightly more complex problem. Suppose we wish to estimate the regression curve that describes the relationship between an outcome variable Y and an exposure variable X, from a sample of data, $\{Y_i, X_i\}$ for $i = 1, ..., n$, drawn independently from the population. This could be done parametrically by, for example, assuming that the relationship is linear and that the errors are normally distributed. We, however, consider the situation in which we do not know which parametric form to specify and hence estimate the regression relationship non-parametrically. In particular, we consider the use of kernel regression. A more comprehensive discussion of non-parametric regression methods can be found elsewhere [30, 34].

Consider the general model

$$Y = g(X) + \epsilon,$$

where ϵ is some error term with $\mathbb{E}[\,\epsilon \,|\, X\,] = 0$. We wish to estimate the function $g(X)$. We can write

$$g(X) = \mathbb{E}[\,Y \,|\, X\,] = \int y \, \frac{f(X,Y)}{f(X)} \, dy.$$

Using kernel density estimation to estimate both the joint density $f(X,Y)$ and the marginal density $f(X)$ gives the Nadarya-Watson estimator [71, 109],

$$\hat{g}(X) = \frac{1}{nh\,\hat{f}(X)} \sum_{i=1}^{n} Y_i \, K\left(\frac{X - X_i}{h}\right), \tag{6.3}$$

where $\hat{f}(X)$ is the Parzen-Rosenblatt estimator given earlier (6.2) and the bandwidth h and kernel function $K(\,\cdot\,)$ are defined as before.

6.2 Kernel density estimation and regression for the propensity score

In order to estimate the variance of the stratified treatment effect estimator from a sample dataset, we will need to estimate both the probability density function of the propensity score and the conditional expectation of the outcome, given the propensity score. We gave a brief overview of kernel density estimation and kernel regression in

Section 6.1.2. We begin by using kernel density estimation to estimate the probability density function of the propensity score. We then estimate the conditional expectation of the outcome given the propensity score using kernel regression. Using these kernel estimators we then estimate the derivatives of the two functions with respect to the propensity score parameters, α.

6.2.1 The kernel density estimator for the propensity score

We now obtain a kernel density estimator for the propensity score. Since, in this chapter, we are assuming that the propensity score is unknown, and therefore must be estimated from the data, we cannot directly observe the propensity scores of our sample. Thus our sample consists of the estimated propensity scores for the sampled subjects, $\{\hat{p}(\mathbf{X}_i)\}$ for $i = 1, ..., n$.

As discussed, we use a normal kernel, defined by $K(u) = e^{-u^2/2}/\sqrt{2\pi}$, in order to obtain a differentiable estimator of the probability density function. We have seen that the optimal choice of bandwidth, h, depends on the variance of the propensity score. We estimate this by the sample variance of the estimated propensity scores, defining

$$\hat{\sigma}_p^2 = \frac{\sum_{i=1}^{n} \left(\hat{p}(\mathbf{X}_i) - \sum_{i=1}^{n} \frac{\hat{p}(\mathbf{X}_i)}{n} \right)^2}{n-1}. \tag{6.4}$$

We set the bandwidth to be $h = 1.06\, n^{-1/5}\, \hat{\sigma}_p$. Then from (6.2), the kernel density estimator for the propensity score is

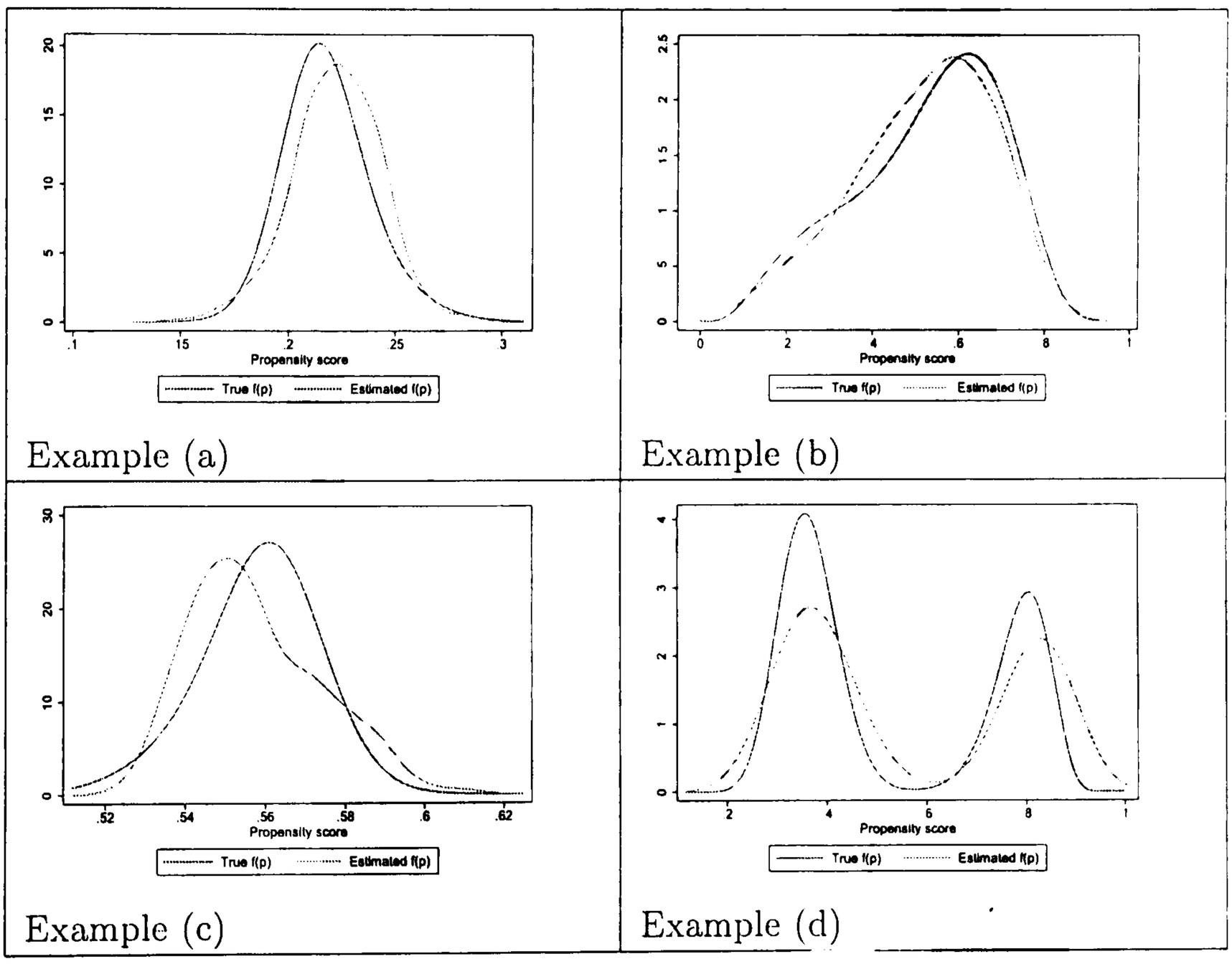
$$\begin{aligned} \hat{f}_p(p) &= \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{p - \hat{p}(\mathbf{X}_i)}{h} \right) \\ &= \frac{1}{nh} \sum_{i=1}^{n} \frac{e^{-\frac{(p-\hat{p}(\mathbf{X}_i))^2}{2h^2}}}{\sqrt{2\pi}}. \end{aligned} \tag{6.5}$$

A sample of 2,000 was simulated from each of the examples (a), (b), (c) and (d) introduced in Sections 5.2.1 – 5.2.4. The density estimator (6.5) was calculated for each of the four datasets. Figure 6.4 shows both the kernel density estimates and the true probability density functions for each of the four examples [1]. The estimators for examples (a), (b) and (c) are fairly close to the true density functions. The kernel

[1] See Appendix C for details of the calculation of the true probability density function of the propensity score.

density estimator for example (d) is less accurate. This is due to the fact that we chose the bandwidth to be optimal for an underlying normal (unimodal) density. In example (d), the underlying density is bimodal which leads to a sample variance that implies that the propensity score is more varied than it actually is. This results in a bandwidth that is too large and so the density is oversmoothed. This problem could be remedied by choosing a smaller bandwidth.

Figure 6.4: *Kernel density estimates for the propensity score, applied to examples (a), (b), (c) and (d) of Chapter 5.*



Derivatives of the kernel density estimator for the propensity score

We now show how the derivative of the probability density function of the propensity score with respect to the propensity score parameters, α_k, for $k = 1, ..., m$,

$$\frac{\partial}{\partial \alpha_k} \{f_p(p; \boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o},$$

can be estimated from a sample dataset. We estimate these derivatives by the derivatives of the kernel density estimator for the propensity score, with respect to the

propensity score parameters, evaluated at $\hat{\boldsymbol{\theta}}$, rather than the unknown $\boldsymbol{\theta}_o$.

From (6.5), the kernel density estimator of the probability density function of the propensity score, viewed as a function of the unknown propensity score parameters, $\boldsymbol{\alpha}$, is

$$\begin{aligned}\hat{f}_p(p;\boldsymbol{\alpha}) &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{p-p\left(\mathbf{X}_i;\boldsymbol{\alpha}\right)}{h}\right)\\ &= \frac{1}{nh}\sum_{i=1}^{n}\frac{e^{-\frac{(p-p\left(\mathbf{X}_i;\boldsymbol{\alpha}\right))^2}{2h^2}}}{\sqrt{2\pi}}.\end{aligned} \tag{6.6}$$

Since changing the parameter α_k affects the density function $f_p(\cdot)$ only through the change in the distribution of the covariates given p, the only part of the kernel density estimate of $f_p(\cdot)$ which depends on α_k is $p\left(\mathbf{X}_i;\boldsymbol{\alpha}\right)$. Differentiating (6.6) with respect to α_k gives

$$\begin{aligned}\frac{\partial \hat{f}_p(p)}{\partial \alpha_k} &= \frac{1}{nh}\sum_{i=1}^{n}\frac{\partial}{\partial p\left(\mathbf{X}_i;\boldsymbol{\alpha}\right)}\left\{K\left(\frac{p-p\left(\mathbf{X}_i;\boldsymbol{\alpha}\right)}{h}\right)\right\}\frac{\partial p\left(\mathbf{X}_i;\boldsymbol{\alpha}\right)}{\partial \alpha_k}\\ &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{p-p\left(\mathbf{X}_i;\boldsymbol{\alpha}\right)}{h}\right)\frac{\left(p-p\left(\mathbf{X}_i;\boldsymbol{\alpha}\right)\right)}{h^2}\frac{\partial p\left(\mathbf{X}_i;\boldsymbol{\alpha}\right)}{\partial \alpha_k}.\end{aligned} \tag{6.7}$$

The propensity score is connected to the propensity score parameters as follows,

$$p\left(\mathbf{X}_i;\boldsymbol{\alpha}\right) = \frac{e^{\boldsymbol{\alpha}^T\mathbf{X}_i}}{1+e^{\boldsymbol{\alpha}^T\mathbf{X}_i}},$$

thus, remembering that X_{ki} is the k^{th} covariate observed on the i^{th} subject,

$$\frac{\partial p\left(\mathbf{X}_i;\boldsymbol{\alpha}\right)}{\partial \alpha_k} = X_{ki}\,p\left(\mathbf{X}_i;\boldsymbol{\alpha}\right)\left(1-p\left(\mathbf{X}_i;\boldsymbol{\alpha}\right)\right).$$

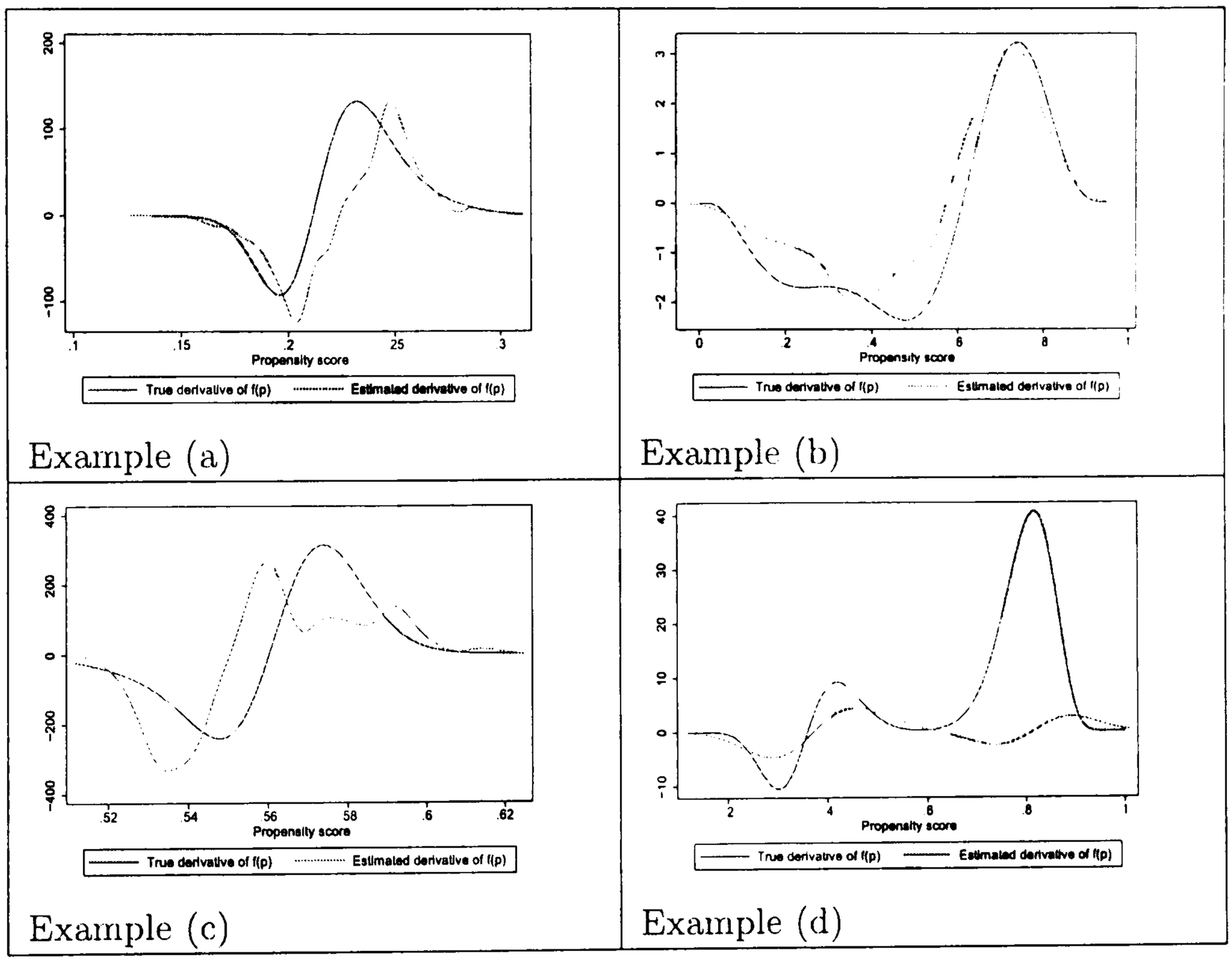
Then substituting these two derivatives into (6.7) and evaluating the derivative at the estimated propensity score parameters, the required derivative estimator is

$$\left.\widehat{\frac{\partial f_p(p)}{\partial \alpha_k}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{p-\hat{p}\left(\mathbf{X}_i\right)}{h}\right)\frac{\left(p-\hat{p}\left(\mathbf{X}_i\right)\right)}{h^2}X_{ki}\,\hat{p}\left(\mathbf{X}_i\right)\left(1-\hat{p}\left(\mathbf{X}_i\right)\right). \tag{6.8}$$

Again, a sample of 2,000 was simulated from each of the examples (a), (b), (c) and (d) introduced in Sections 5.2.1 – 5.2.4. The estimator of the derivative of the probability

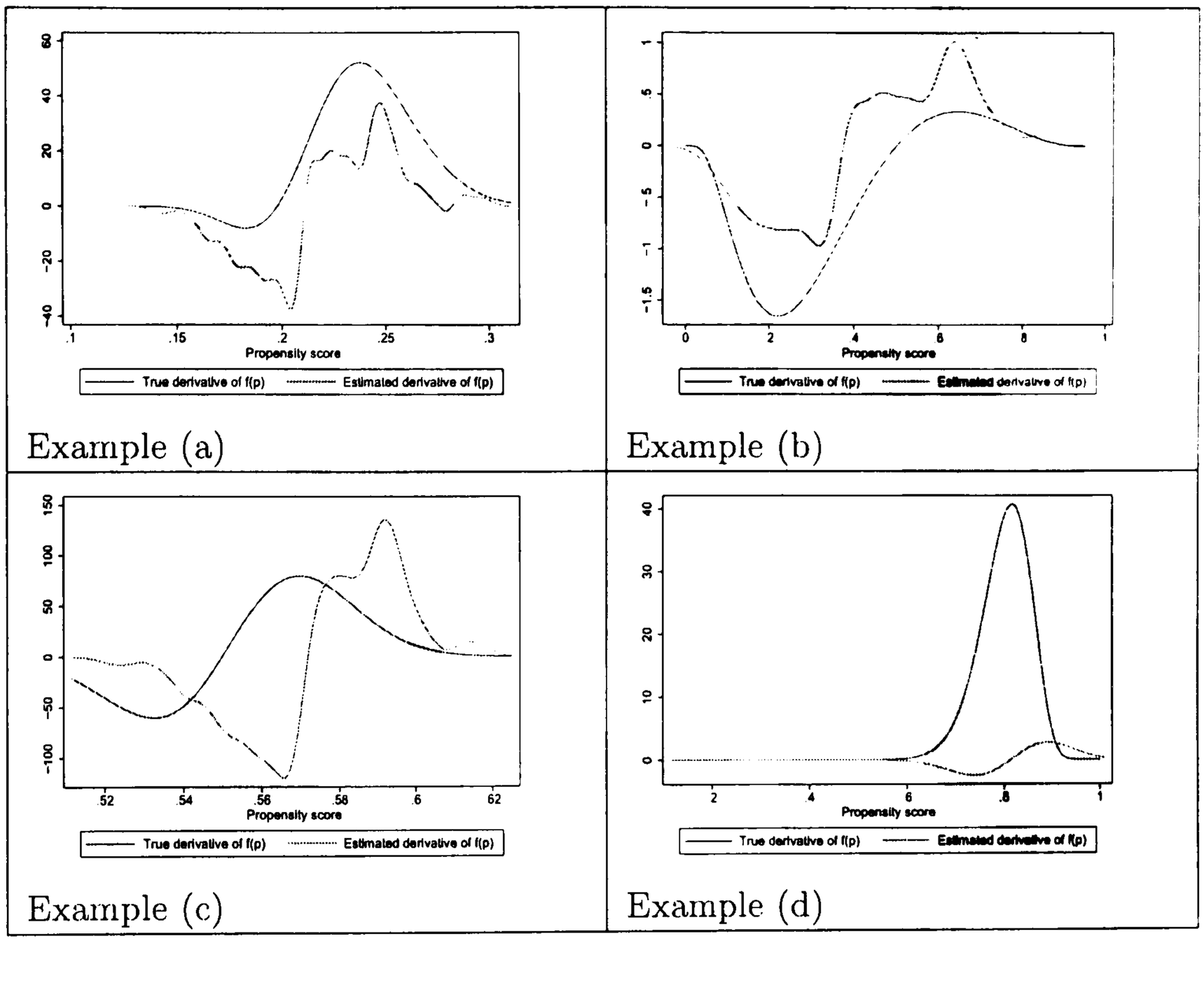
density function with respect to each of the three propensity score parameters, α_0, α_1 and α_2, was calculated for each of them using (6.8). Figure 6.5 shows both the kernel density estimates of the derivatives and the true derivatives, with respect to α_o for each of the four examples. Figure 6.6 shows the same derivatives, with respect to α_1, and Figure 6.7 shows the derivatives with respect to α_2.

Figure 6.5: *Estimated derivatives of the probability density function of the propensity score with respect to α_0, applied to examples (a), (b), (c) and (d) of Chapter 5.*



For examples (a), (b) and (c), the kernel estimates of the derivatives are fairly good approximations, although, as expected, they are less accurate than the kernel density estimates of the probability density functions themselves. The estimated derivatives for example (d) are much smoother than the true derivatives. This is because, as discussed previously, since the true distribution of the propensity score is bimodal, the bandwidth that we used, $h = 1.06\, n^{-1/5}\, \hat{\sigma}_p$, is too large and the kernel density estimate over-smooths the data. This implies that the magnitude of the estimated derivatives will be too small, which is exactly what we see in Figures 6.5 – 6.7.

Figure 6.6: *Estimated derivatives of the probability density function of the propensity score with respect to α_1, applied to examples (a), (b), (c) and (d) of Chapter 5.*



6.2.2 *The kernel regression of the outcome on the propensity score*

We now turn to kernel regression to estimate the conditional expectation of the potential outcomes, (Y_0, Y_1), given the propensity score,

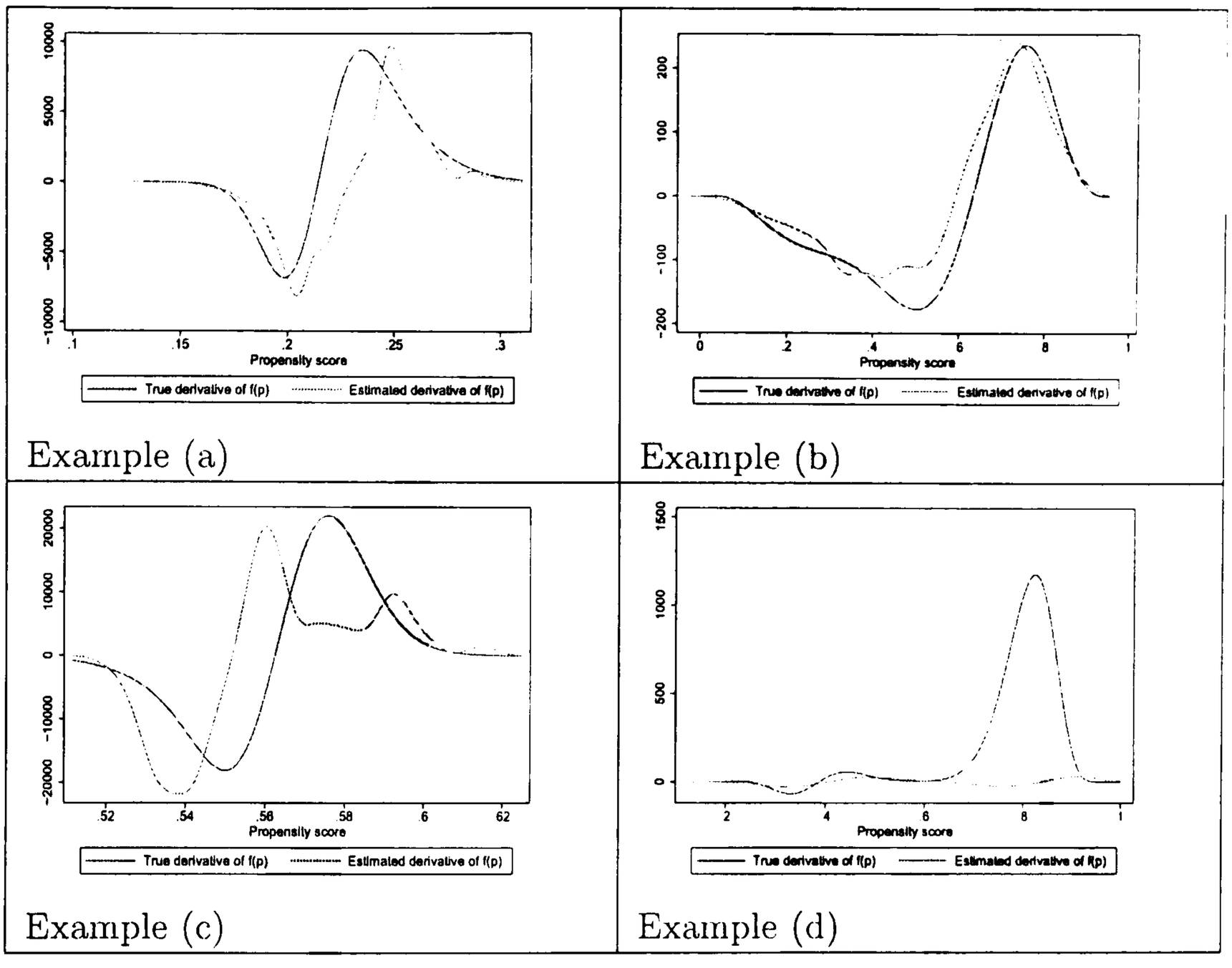
$$\mathbb{E}\left[\, Y_1 \,|\, p\left(\mathbf{X}\right) = p \,\right], \qquad \mathbb{E}\left[\, Y_0 \,|\, p\left(\mathbf{X}\right) = p \,\right].$$

We show how the first of these two expectations is estimated. To estimate the conditional expectation, we need the potential outcome Y_1 rather than the observed outcome, Y, which depends on the treatment status. We could re-write the expectation in terms of the observed outcome and treatment status, since

$$\mathbb{E}\left[\, Y_1 \,|\, p\left(\mathbf{X}\right) = p \,\right] = \frac{\mathbb{E}\left[\, Y\, Z \,|\, p\left(\mathbf{X}\right) = p \,\right]}{\mathbb{P}(Z = 1 \,|\, p\left(\mathbf{X}\right) = p)}.$$

We could then use kernel estimation techniques to estimate the expression on the right-hand side of this equation, requiring only the observed outcome rather than the

Figure 6.7: *Estimated derivatives of the probability density function of the propensity score with respect to α_2, applied to examples (a), (b), (c) and (d) of Chapter 5.*



potential outcomes. We found, however, that more accurate results could be obtained by estimating

$$\hat{Y}_{1i} = \begin{cases} Y_i & \text{if } Z_i = 1 \\ Y_i + \hat{\beta}^s & \text{if } Z_i = 0. \end{cases}$$

Then our sample consists of the estimated potential outcomes and propensity scores $\{\hat{Y}_{1i}, \hat{p}\,(\mathbf{X}_i)\}$ for $i = 1, ..., n$. As before, we use a normal kernel, defined by $K(u) = e^{-u^2/2}/\sqrt{2\pi}$, in order to obtain a differentiable estimator of the conditional expectation. We define $h = 1.06\, n^{-1/5}\, \hat{\sigma}_p$ where $\hat{\sigma}_p^2$ is the sample variance of the estimated propensity score (6.4). Then the Nadarya-Watson estimator (6.3) for the conditional expectation of Y_1 given the propensity score is

$$\hat{\mathbb{E}}\left[\, Y_1 \,|\, p\,(\mathbf{X}) = p \,\right] = \frac{1}{nh\, \hat{f}_{\boldsymbol{p}}(p)} \sum_{i=1}^{n} \hat{Y}_{1i}\, K\left(\frac{p - \hat{p}(\mathbf{X}_i)}{h}\right), \tag{6.9}$$

where $\hat{f}_p(p)$ is the kernel density estimator for the propensity score derived earlier (6.5).

Derivatives of the kernel regression of the outcome on the propensity score

We now show how the derivatives of the conditional expectations of the potential outcomes, (Y_0, Y_1), given the propensity score, with respect to the propensity score parameters, α_k, for $k = 1, ..., m$,

$$\frac{\partial}{\partial \alpha_k} \{\mathbb{E}[Y_1 \,|\, p(\mathbf{X}) = p]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}, \qquad \frac{\partial}{\partial \alpha_k} \{\mathbb{E}[Y_0 \,|\, p(\mathbf{X}) = p]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o},$$

can be estimated from a sample dataset. We estimate these derivatives by the derivatives of the kernel regression estimator of the expected outcome given the propensity score, with respect to the propensity score parameters, evaluated at $\hat{\boldsymbol{\theta}}$, rather than the unknown $\boldsymbol{\theta}_o$.

From (6.9), the kernel regression estimator of the conditional outcome Y_1 given the propensity score, viewed as a function of the unknown propensity score parameters, $\boldsymbol{\alpha}$, is

$$\hat{\mathbb{E}}[Y_1 \,|\, p(\mathbf{X}; \boldsymbol{\alpha}) = p] = \frac{1}{nh\, \hat{f}_p(p; \boldsymbol{\alpha})} \sum_{i=1}^{n} \hat{Y}_{1i}\, K\left(\frac{p - p(\mathbf{X}_i; \boldsymbol{\alpha})}{h}\right), \tag{6.10}$$

The only parts of this which depend on α_k are $p(\mathbf{X}_i; \boldsymbol{\alpha})$ and the kernel density estimator $\hat{f}_p(p; \boldsymbol{\alpha})$. Differentiating (6.10) with respect to α_k using the quotient rule and evaluating the resulting expression at the estimated propensity score parameters gives the required derivative estimator

$$\begin{aligned}\left.\frac{\partial \mathbb{E}[Y_1 \,|\, \widehat{p(\mathbf{X}; \boldsymbol{\alpha})} = p]}{\partial \alpha_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} &= -\frac{1}{nh\hat{f}_p(p)^2} \sum_{i=1}^{n} \hat{Y}_{1i}\, K\left(\frac{p - \hat{p}(\mathbf{X}_i)}{h}\right) \left.\frac{\widehat{\partial f_p(p; \boldsymbol{\alpha})}}{\partial \alpha_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \\ &+ \frac{1}{\hat{f}_p(p)nh} \sum_{i=1}^{n} \hat{Y}_{1i}\, K\left(\frac{p - \hat{p}(\mathbf{X}_i)}{h}\right) \frac{(p - \hat{p}(\mathbf{X}_i))}{h^2} X_{ki}\, \hat{p}(\mathbf{X}_i)(1 - \hat{p}(\mathbf{X}_i)).\end{aligned} \tag{6.11}$$

6.3 Estimating the four variance components from a sample dataset

We now consider the estimation of the variance of the stratified treatment effect estimator from a sample dataset, when the propensity score is estimated using a

correctly specified logistic regression model. This variance, denoted by $\mathbb{V}_e[\hat{\beta}^s]$, can be expressed as the sum of four variance components,

$$\mathbb{V}_e[\hat{\beta}^s] = V_1 + V_2 + V_3 + V_4,$$

where the four variance components, V_1, V_2, V_3 and V_4, are defined in Theorems 3.1 and 3.2. We now consider the estimation of each of the four variance components in turn.

6.3.1 Estimating the variance component V_1

From Theorem 3.1, the first variance component is defined as

$$V_1 = \frac{1}{n}\sum_{s=1}^{K} r_s^2 \left\{ \frac{\mathbb{V}[Y \mid Z=1, S_{so}=1]}{d_{so}} + \frac{\mathbb{V}[Y \mid Z=0, S_{so}=1]}{r_s - d_{so}} \right\}.$$

The r_s are fixed and known. We therefore only need to estimate the population probabilities of being treated and in each stratum, $\mathbf{d}_o$, and the within-stratum outcome variances. For $s = 1, ..., K$,

$$d_{so} = \mathbb{E}[Z\, S_{so}] = \mathbb{P}(Z=1,\, S_{so}=1)$$

which can be estimated by the proportion of the sample who are treated and in that sample stratum,

$$\hat{d}_{so} = \frac{1}{n}\sum_{i=1}^{n} Z_i\, \hat{S}_{si},$$

where $\hat{S}_{si} = 1_{[\hat{q}_{s-1} \le \hat{p}(\mathbf{X}) < \hat{q}_s]}$ is an indicator for the s^{th} sample stratum. We can estimate the variance of the outcome, conditional on being treated and in the s^{th} population stratum by the sample variance of the outcomes of those subjects who are treated and in the s^{th} sample stratum,

$$\hat{\mathbb{V}}[Y \mid Z=1, S_{so}=1] = \frac{\sum_{i=1}^{n}\left(Y_i\, Z_i\, \hat{S}_{si} - \frac{\sum_{i=1}^{n} Y_i\, Z_i\, \hat{S}_{si}}{\sum_{i=1}^{n} Z_i\, \hat{S}_{si}}\right)^2}{\sum_{i=1}^{n} Z_i\, \hat{S}_{si} - 1},$$

and similarly,

$$\hat{\mathbb{V}}[Y \mid Z=0, S_{so}=1] = \frac{\sum_{i=1}^{n}\left(Y_i\,(1-Z_i)\, \hat{S}_{si} - \frac{\sum_{i=1}^{n} Y_i\,(1-Z_i)\, \hat{S}_{si}}{\sum_{i=1}^{n} (1-Z_i)\, \hat{S}_{si}}\right)^2}{\sum_{i=1}^{n} (1-Z_i)\, \hat{S}_{si} - 1}.$$

6.3.2 Estimating the variance component V_2

As we discussed in Section 3.4.2, we expect the component V_2 to be negligible and so it may not be necessary to estimate it. However, we now show how it can be estimated from a dataset, if desired. From Theorem 3.1, the second variance component is defined as

$$V_2 = \frac{1}{n} \left.\frac{\partial \beta^*}{\partial \mathbf{q}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} (n\mathrm{Cov}[\,\hat{\mathbf{q}}\,]) \left.\frac{\partial \beta^*}{\partial \mathbf{q}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

Estimating $(n\,\mathbf{Cov}[\hat{\mathbf{q}}])$

We begin with the covariance matrix $(n\,\mathrm{Cov}[\hat{\mathbf{q}}])$, defined, for $j, k = 1, ..., K-1,\ j \geq k$, as

$$(n\ \mathrm{Cov}[\,\hat{\mathbf{q}}\,])_{jk} = \frac{\mathbb{P}(p_o(\mathbf{X}) > q_{jo})\ \mathbb{P}(p_o(\mathbf{X}) < q_{ko})}{f_p(q_{jo})\ f_p(q_{ko})}.$$

The probability that the propensity score is less than (or greater than) the s^{th} population strata boundary can be estimated by the proportion of the estimated propensity scores in the sample that are less than (or greater than) the s^{th} estimated strata boundary, so for example,

$$\hat{\mathbb{P}}(p_o(\mathbf{X}) < q_{2o}) = \frac{1}{n} \sum_{i=1}^{n} 1_{[\hat{p}\,(\mathbf{X}) < \hat{q}_2\,]}.$$

We must also estimate $f_p(q_{jo})$, the probability density function of the propensity score at the j^{th} population strata boundary, for $j = 1, .., K-1$. We have already shown how to construct a kernel density estimator for the propensity score (Section 6.2.1). We merely evaluate this estimator of the probability density function of the propensity score at the sample strata boundaries. Then, defining $h = 1.06\ n^{-1/5}\,\hat{\sigma}_p$, where $\hat{\sigma}_p^2$ is the sample variance of the estimated propensity score, the required probability density function estimate is

$$\hat{f}_p(q_{jo}) = \frac{1}{nh} \sum_{i=1}^{n} \frac{e^{-\frac{(\hat{q}_j - \hat{p}\,(\mathbf{X}_i))^2}{2h^2}}}{\sqrt{2\pi}}, \tag{6.12}$$

Estimating $\left.\frac{\partial \beta^*}{\partial \mathbf{q}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$

This derivative is defined as

$$\left.\frac{\partial \beta^*}{\partial \mathbf{q}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \frac{\partial}{\partial \mathbf{q}^T}\left\{\sum_{s=1}^{K} r_s\{\mathbb{E}[Y \mid Z=1,\, S_s=1] - \mathbb{E}[Y \mid Z=0,\, S_s=1]\}\right\}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

We now show how we estimate $\frac{\partial}{\partial q_1}\{\mathbb{E}[Y \mid Z=1,\, S_1=1]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$. The other elements of the derivative can be estimated in a similar fashion.

Since

$$\begin{aligned} f(p \mid Z=1,\, S_1=1) &= \frac{\mathbb{P}(Z=1,\, S_1=1 \mid p_o(\mathbf{X};\boldsymbol{\alpha})=p)\, f_p(p)}{\mathbb{P}(Z=1,\, S_1=1)} \\ &= \frac{p\, f_p(p)\, 1_{[p \in S_1]}}{\mathbb{P}(Z=1,\, S_1=1)}, \end{aligned}$$

and

$$\mathbb{P}(Z=1,\, S_1=1) = \int_0^{q_1} p\, f_p(p)\, dp,$$

we can write

$$\mathbb{E}[Y \mid Z=1,\, S_1=1] = \frac{\int_0^{q_1} \mathbb{E}[Y \mid Z=1,\, p_o(\mathbf{X})=p]\, p\, f_p(p)\, dp}{\int_0^{q1} p\, f_p(p)\, dp}. \tag{6.13}$$

We now appeal to the fundamental theorem of calculus, since we have already assumed that both the probability density function of the propensity score and the conditional expectation of the outcome given the propensity score are continuous in p. Remembering that when evaluated at the true strata boundaries the denominator of (6.13) is equal to d_{so}, differentiating (6.13) using the quotient rule, with respect to q_1, and evaluating at the population strata boundaries gives

$$\begin{aligned} \left.\frac{\partial}{\partial q_1}\{\mathbb{E}[Y \mid Z=1,\, S_1=1]\}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} &= \frac{\mathbb{E}[Y \mid Z=1,\, p_o(\mathbf{X})=q_1]\, q_1\, f_p(q_1)}{d_{1o}} \\ &\quad - \frac{\mathbb{E}[Y \mid Z=1,\, S_{1o}=1]\, q_1\, f_p(q_1)}{d_{1o}}. \end{aligned}$$

Due to the balancing property of the propensity score (Chapter 2),

$$\mathbb{E}\,[\,Y \mid Z = 1,\, p_o(\mathbf{X}) = p\,] = \mathbb{E}\,[\,Y_1 \mid p_o(\mathbf{X}) = p\,].$$

So

$$\frac{\partial}{\partial q_1}\{\mathbb{E}\,[Y|Z=1, S_1=1]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = q_{1o} f_p(q_{1o}) \frac{\{\mathbb{E}\,[\,Y_1 \mid p_o(\mathbf{X}) = q_{1o}\,] - \mathbb{E}\,[Y|Z=1, S_{1o}=1]\}}{d_{1o}}.$$

The quantities on the right-hand side of this equation are easily estimated from the data. We estimate the population strata boundary by its sample estimate. We have already shown how the probability density function of the propensity score, $f_p(q_{1o})$, can be estimated using kernel density estimation (6.12). We use kernel regression to estimate the conditional expectation of the outcome, given the propensity score. The remaining two quantities, $\mathbb{E}\,[\,Y \mid Z = 1,\, S_{1o} = 1\,]$ and d_{1o}, can be estimated using sample averages. However, we have found that the whole derivative is much more precisely estimated if we also use kernel density estimates of these two quantities. We can write

$$d_{1o} = \int_0^{q_1} p\, f_p(p)\, dp.$$

To estimate this integral, we could replace the population strata boundary by its sample estimate, $\hat{q}_1$, replace the probability density function, $f_p(\,\cdot\,)$, by its kernel density estimate (6.5), and then use the trapezium rule (Section 6.1.1) to estimate the integral. Suppose we partition the interval $[\,0, \hat{q}_1\,]$ into 50 equal width sub-intervals, with $0 = p_0 < p_1 < \ldots < p_{50} = \hat{q}_1$. Then

$$\hat{d}_1 = \frac{\hat{q}_1}{50}\{p_1\, \hat{f}_p(p_1) + \ldots + p_{50}\, \hat{f}_p(p_{50})\} - \frac{\hat{q}_1}{2 \times 50}\,\{\,p_{50}\, \hat{f}_p(p_{50})\,\}.$$

Similarly, using the definition of the conditional outcome given treatment and strata calculated previously (6.13), we can estimate $\mathbb{E}\,[\,Y \mid Z = 1,\, S_{1o} = 1\,]$ using the trapezium rule and the kernel regression estimate of the conditional outcome given the propensity score.

6.3.3 Estimating the variance component V_3

From Theorem 3.2, the third variance component is defined as

$$V_3 = -\frac{1}{n}\,\mathbf{C}\,(n\,\mathrm{Cov}[\,\hat{\boldsymbol{\alpha}}\,])\,\mathbf{C}^T.$$

Estimating $(n\ \mathbf{Cov}[\hat{\boldsymbol{\alpha}}])$

The covariance matrix $(n\,\mathrm{Cov}[\,\hat{\boldsymbol{\alpha}}\,])$ is defined in terms of its inverse, for $j, k = 1, ..., m$, as

$$(n\ \mathrm{Cov}[\,\hat{\boldsymbol{\alpha}}\,])^{-1}_{jk} = \mathbb{E}\,[\,p_o(\mathbf{X})\,(1 - p_o(\mathbf{X}))\ X_j\ X_k\,].$$

The covariance matrix above is easily estimated by substituting a sample average for the expectation,

$$\hat{\mathbb{E}}\,[\,p_o(\mathbf{X})\,(1 - p_o(\mathbf{X}))\ X_j\ X_k\,] = \sum_{i=1}^{n} \hat{p}\,(\mathbf{X}_i)\,(1 - \hat{p}\,(\mathbf{X}_i))X_{ji}X_{ki}.$$

Estimating C

The term $\mathbf{C} = (C_1, ..., C_m)$ is defined as

$$\begin{aligned} C_k = & \sum_{s=1}^{K} r_s\ \mathrm{Cov}[\,Y,\ X_k\,(1 - p_o(\mathbf{X}))\,|\,Z = 1,\ S_{so} = 1\,] \\ + & \sum_{s=1}^{K} r_s\ \mathrm{Cov}[\,Y,\ X_k\,p_o(\mathbf{X})\,|\,Z = 0,\ S_{so} = 1\,]. \end{aligned}$$

We show how to estimate the covariance $\mathrm{Cov}[\,Y,\ X_k\,(1 - p_o(\mathbf{X}))\,|\,Z = 1,\ S_{1o} = 1\,]$. All the other covariances can be estimated in a similar manner. Using the equation linking correlation and covariance, $\mathrm{Cov}[\,A,\ B\,] = \mathrm{Corr}[\,A,\ B\,]\sqrt{\mathbb{V}\,[\,A\,]}\sqrt{\mathbb{V}\,[\,B\,]}$, we have

$$\begin{aligned} & \mathrm{Cov}[\,Y,\ X_k\,(1 - p_o(\mathbf{X}))\,|\,Z = 1,\ S_{1o} = 1\,] \\ = & \mathrm{Corr}[\,Y,\ X_k\,(1 - p_o(\mathbf{X}))\,|\,Z = 1,\ S_{1o} = 1\,]\,\sigma_Y\,\sigma_{Xp}, \end{aligned}$$

where σ_Y is the variance of Y conditional on being treated and in the first population stratum, and σ_{Xp} is the variance of $X_k\,(1 - p_o(\mathbf{X}))$ conditional on being treated and in the first population stratum. We then merely estimate the two variances and the correlation by their sample estimates.

6.3.4 Estimating the variance component V_4

From Theorem 3.2, the fourth variance component is defined as

$$V_4 = \frac{1}{n}\ \mathbf{e}\,(n\,\mathrm{Cov}[\,\hat{\boldsymbol{\alpha}}\,])\ \mathbf{e}^T.$$

where, $\mathbf{e} = (e_1, ..., e_m)$, for $k = 1, ..., m$, is defined as $e_k = e_{\boldsymbol{\alpha}k} + e_{\mathbf{q}k}$.

Estimating $e_{\mathbf{q}k}$

For $k = 1, ..., m$,

$$e_{\mathbf{q}k} = \sum_{j=1}^{K-1} \left.\frac{\partial \beta^*}{\partial q_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} f_p(q_{jo})^{-1}\ \frac{\partial}{\partial \alpha_k}\ \{\mathbb{E}\,[\,1_{[p\,(\mathbf{X};\boldsymbol{\alpha})<q_j]}\,]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

We showed, in Section 6.3.2 how we could use kernel density estimation to estimate both the derivative $\left.\frac{\partial \beta^*}{\partial q_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$ and the probability density function of the propensity score, $f_p(q_{jo})$. We now discuss how to estimate

$$\frac{\partial}{\partial \alpha_k}\ \{\mathbb{E}\,[\,1_{[p\,(\mathbf{X};\boldsymbol{\alpha})<q_j]}\,]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

We can write this as

$$\frac{\partial}{\partial \alpha_k}\ \{\mathbb{E}\,[\,1_{[p\,(\mathbf{X};\boldsymbol{\alpha})<q_j]}\,]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \left.\frac{\partial}{\partial \alpha_k}\int_0^{q_{jo}} f(p;\boldsymbol{\alpha})\,dp\ \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o.}$$

Interchanging the order of differentiation and integration gives

$$\frac{\partial}{\partial \alpha_k}\ \{\mathbb{E}\,[\,1_{[p\,(\mathbf{X};\boldsymbol{\alpha})<q_j]}\,]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \int_0^{q_{jo}} \frac{\partial}{\partial \alpha_k}\ \{f(p;\boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\ dp.$$

To estimate this integral, we merely substitute the kernel estimate of the derivative of the probability density function of the propensity score (6.8), with respect to α_k, and estimate the integral using the trapezium rule (Section 6.1.1).

Estimating $e_{\alpha k}$

For $k = 1, ..., m$,

$$e_{\alpha k} = \sum_{s=1}^{K} r_s \left\{ \frac{(I_{Y_1 k} - \mathbb{E}[Y \mid Z = 1, S_{so} = 1] I_{f_1 k})}{d_{so}} - \frac{(I_{Y_0 k} - \mathbb{E}[Y \mid Z = 0, S_{so} = 1] I_{f_0 k})}{r_s - d_{so}} \right\}.$$

As for the component V_2, although we could estimate $\mathbb{E}[Y \mid Z = 1, S_{so} = 1]$ and d_{so} by a simple sample average, we estimate them using kernel density estimation techniques (see Section 6.3.2) since this appears to produce more precise estimates of V_4. Now

$$I_{f_1 k} = \int_{q_{(s-1)o}}^{q_{so}} r \frac{\partial}{\partial \alpha_k} \{f_p(r; \boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \, dr.$$

Then $I_{f_1 k}$ can be estimated very simply by replacing the derivative of the probability density function of the propensity score with its kernel density estimate (6.8) and using the trapezium rule (Section 6.1.1) to estimate the integral. We estimate $I_{f_0 k}$ in the same way.

The integral $I_{Y_1 k}$ is defined as

$$I_{Y_1 k} = \int_{q_{(s-1)o}}^{q_{so}} r \frac{\partial}{\partial \alpha_k} \{\mathbb{E}[Y_1 \mid p(\mathbf{X}; \boldsymbol{\alpha}) = r] f_p(r; \boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \, dr.$$

We can write

$$\begin{aligned} I_{Y_1 k} = & \int_{q_{(s-1)o}}^{q_{so}} r \, \mathbb{E}[Y_1 \mid p(\mathbf{X}; \boldsymbol{\alpha}) = r] \frac{\partial}{\partial \alpha_k} \{f_p(r; \boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \, dr \\ & + \int_{q_{(s-1)o}}^{q_{so}} r \, f_p(r; \boldsymbol{\alpha}) \frac{\partial}{\partial \alpha_k} \{\mathbb{E}[Y_1 \mid p(\mathbf{X}; \boldsymbol{\alpha}) = r]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \, dr. \end{aligned}$$

We have already seen how we can obtain kernel density estimates of the two derivatives above, the conditional expectation and the probability density function of the propensity score. These are substituted into the equation above and the trapezium rule is used to estimate the integral.

6.4 An alternative approach

Recollect that we calculated the variances $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$ using the formula

$$\mathbb{V}[\,\hat{\beta}^s\,] = \mathbf{A^{-1}BA^{-T}}, \tag{6.14}$$

with

$$\mathbf{A} = \mathbb{E}\left[-\frac{\partial}{\partial\boldsymbol{\theta}^T}\{\boldsymbol{\psi}(\mathbf{W};\,\boldsymbol{\theta})\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\right],$$

and

$$\mathbf{B} = \mathbb{E}\,[\,\boldsymbol{\psi}(\mathbf{W};\,\boldsymbol{\theta}_o)\;\boldsymbol{\psi}^T(\mathbf{W};\,\boldsymbol{\theta}_o)\,],$$

where the function $\boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta})$ is defined in Sections 3.2 and 3.3 for the situations where the propensity score is known and estimated, respectively. By calculating all the components of the matrices A and B analytically and multiplying out the matrices we obtained the formulæ $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$. However, there is an alternative — and much simpler — method of obtaining an estimator of the variance of the stratified treatment effect estimator. We simply estimate each component of the matrices A and B by a sample estimate and then substitute the estimates of A and B into (6.14). This is then an estimate of the variance of the stratified treatment effect estimator.

To obtain a sample estimate of the matrix B we simply replace the expectation by a sample average. So

$$\hat{\mathbf{B}}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}_j(\mathbf{W}_i;\,\hat{\boldsymbol{\theta}})\,\boldsymbol{\psi}_k(\mathbf{W}_i;\,\hat{\boldsymbol{\theta}}).$$

Similarly, when the component $\boldsymbol{\psi}_j$ is differentiable with respect to $\boldsymbol{\theta}_k$, we can estimate

$$\hat{\mathbf{A}}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left[-\frac{\partial}{\partial\boldsymbol{\theta}_k}\{\boldsymbol{\psi}_j(\mathbf{W}_i;\,\boldsymbol{\theta})\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]. \tag{6.15}$$

When, however, the component $\boldsymbol{\psi}_j$ is not differentiable with respect to $\boldsymbol{\theta}_k$, this is not possible. For example, we cannot estimate the sub-matrix a_{14} in this way (Appendix B). In this case, we use the alternative definition of the matrix A,

$$\mathbf{A} = \frac{\partial}{\partial\boldsymbol{\theta}^T}\left\{\mathbb{E}\,[-\boldsymbol{\psi}(\mathbf{W}\,;\,\boldsymbol{\theta})]\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}, \tag{6.16}$$

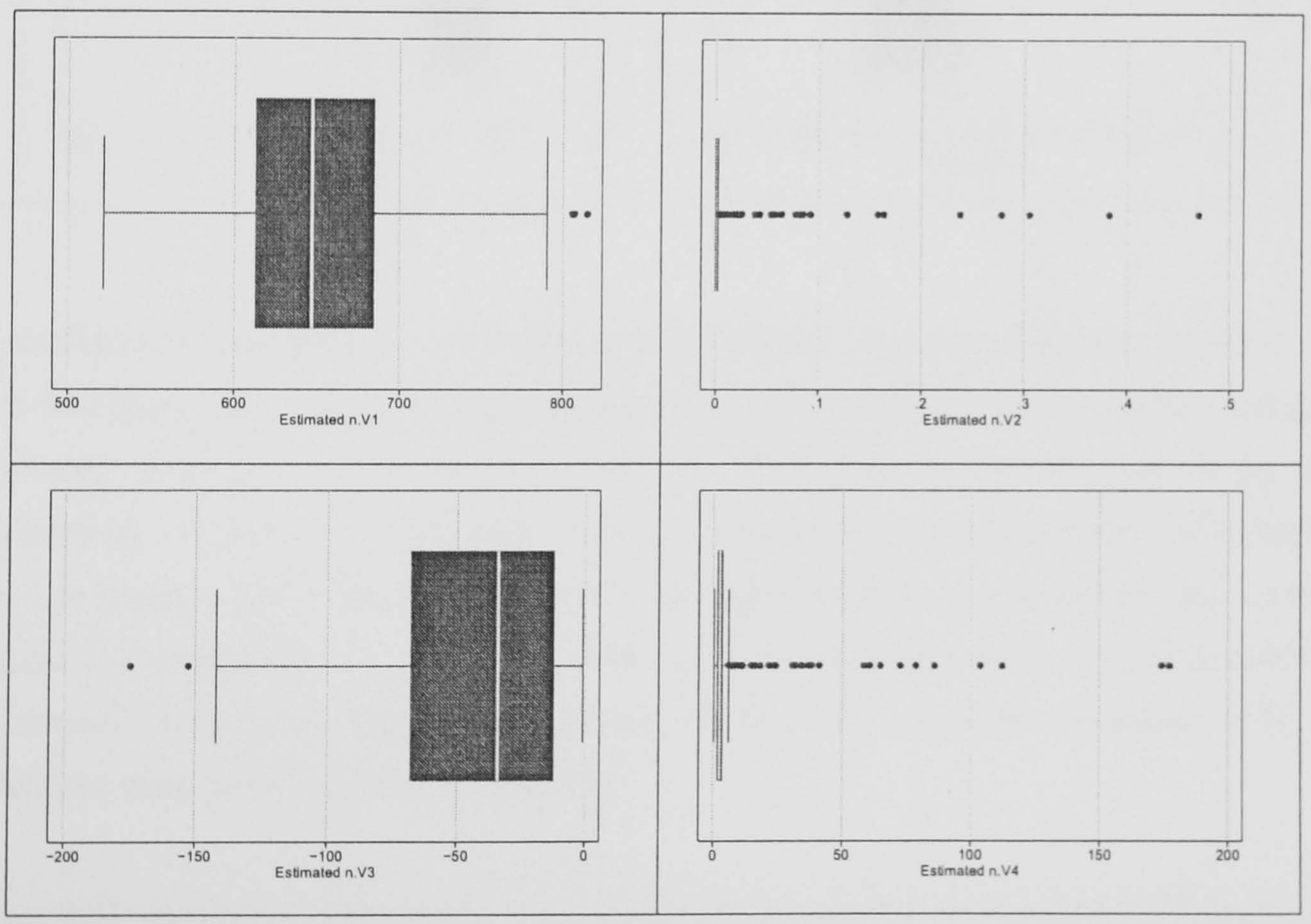
which we then estimate using the kernel density methods described earlier in this chapter.

We describe this method as the 'direct' method, and the variance formulæ calculated earlier as the 'components' method.

6.5 Estimating the variance using hypothetical examples

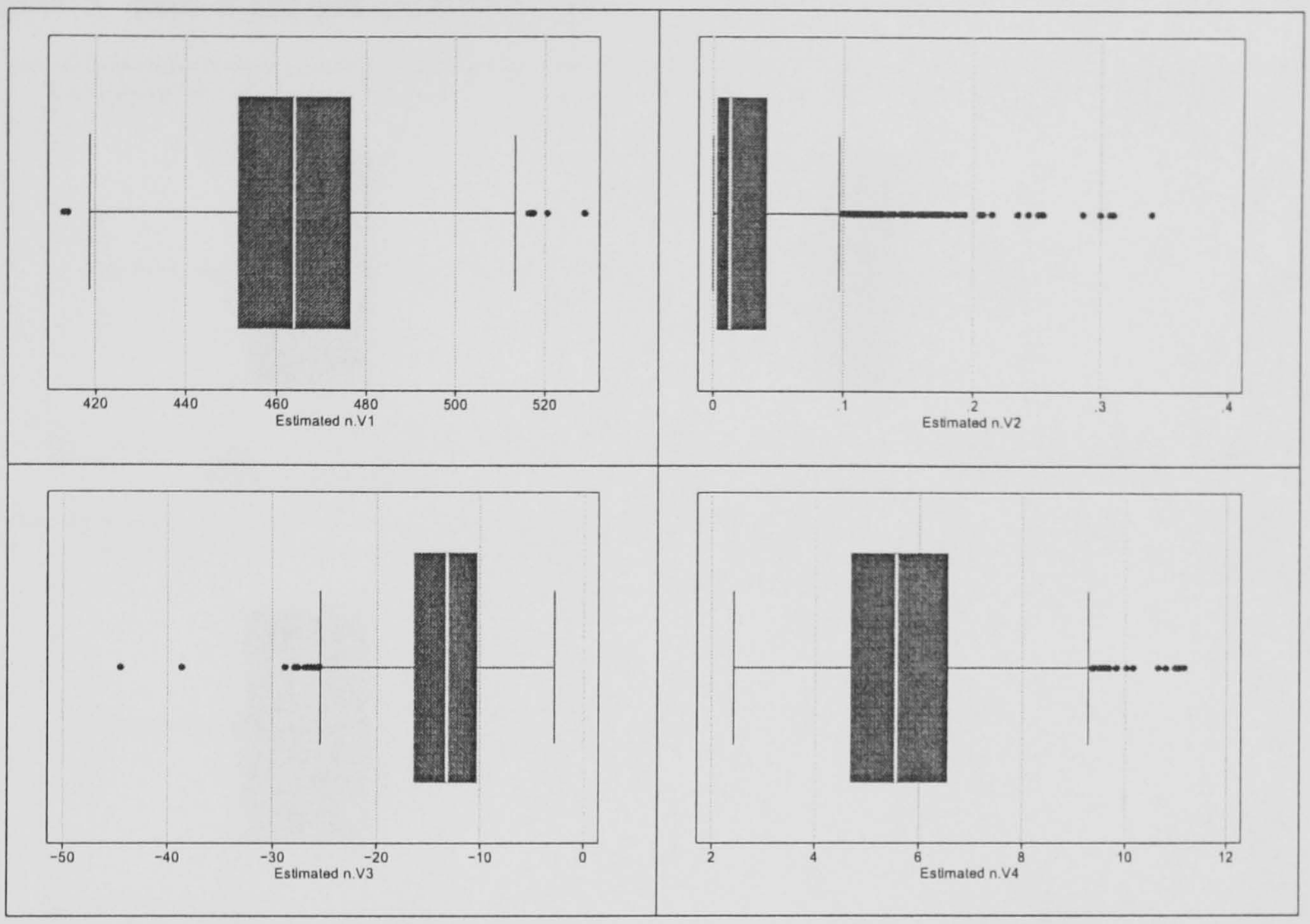
In order to see how precisely we can estimate the four variance components from a sample dataset, we simulated 1,000 datasets each of size $n = 2,000$ from examples (a) and (b) and estimated the four variance components for each dataset. Figure 6.8 shows boxplots of the resulting 1,000 estimated variance components for example (a) and Figure 6.9 shows the analogous boxplots for example (b).

Figure 6.8: *Boxplots showing the range of estimates of the four variance components from 1,000 simulated datasets for hypothetical example (a) of Chapter 5, with a sample size of 2,000.*



We see that the estimated variance components V_1 and V_3 are very variable. This is because we are not making any parametric assumptions about the variance of the outcome and hence we may only have four or five observations with which to estimate a within-stratum outcome variance. The two components V_1 and V_3 are, however, negatively correlated since they both depend on the within-strata outcome variance. Thus some of the error incurred in estimating these components is cancelled out when they are added together.

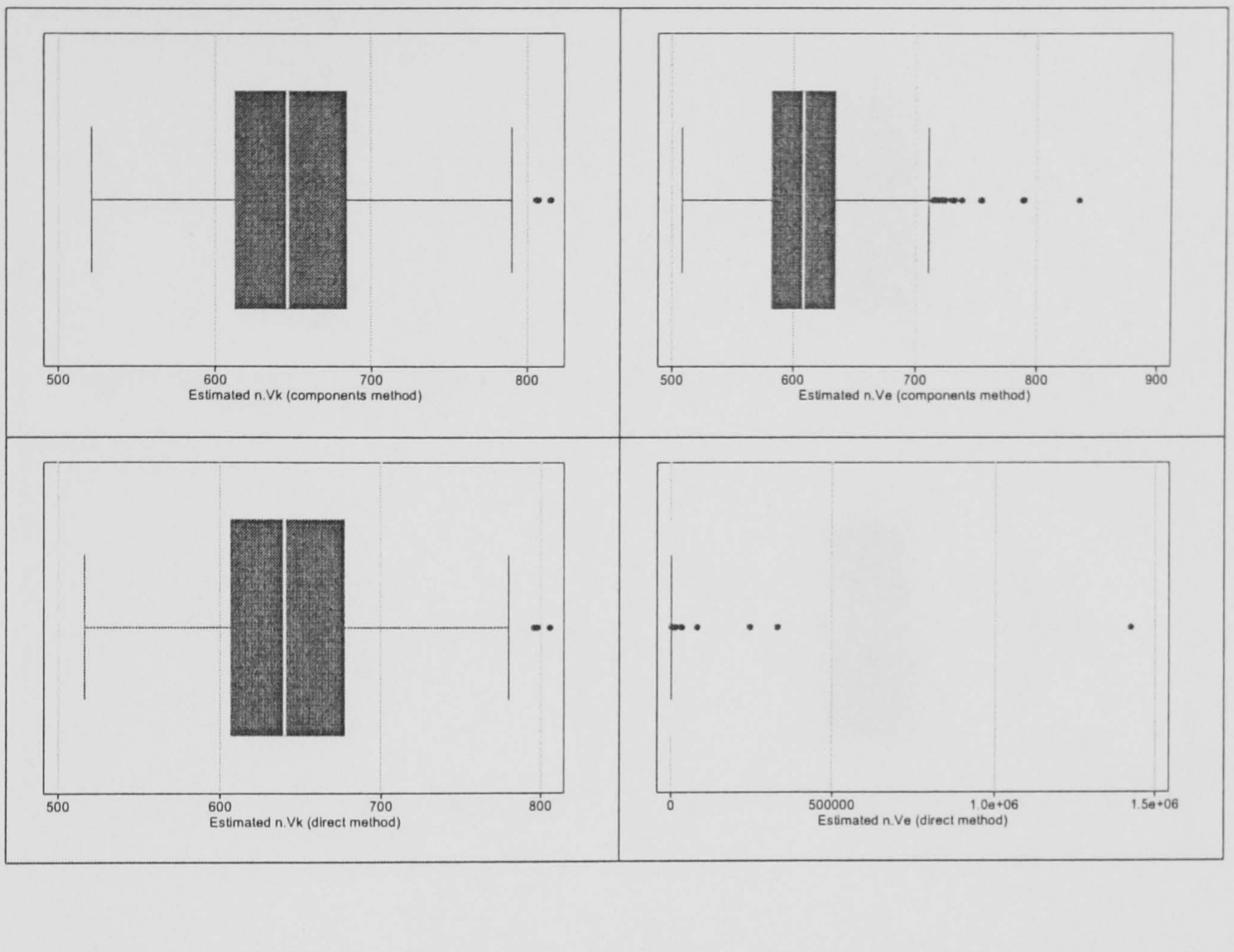
Figure 6.9: *Boxplots showing the range of estimates of the four variance components from 1,000 simulated datasets for hypothetical example (b) of Chapter 5, with a sample size of 2,000.*



The variance component V_4 is mostly well estimated. In some datasets, however, it is much too large. This tends to happen when the empirical distribution of the estimated propensity score in the sample dataset is very dissimilar to the true propensity score distribution. However, in practice we do not know when this occurs. Although we have not proved that the sum $V_3 + V_4$ is always negative, a procedure that uses the estimated variance $V_1 + V_2 + V_3 + V_4$ whenever the estimate of $V_3 + V_4$ is negative, as we expect it to be, but uses the estimated variance $V_1 + V_2$ if the estimate of $V_3 + V_4$ is positive may perform well in practice.

Figure 6.10 shows boxplots of the estimated variances $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$ from 1,000 simulated datasets for hypothetical example (a) of Chapter 5, each of size $n = 2,000$. These two variances were estimated using two different methods. Firstly, each variance was estimated by separately estimating the four variance components V_1, V_2, V_3 and V_4 (the components method). Secondly, the two variances were estimated by the direct method (Section 6.4). Figure shows the boxplots for example (b). We see that although the variances are mostly well estimated, in some datasets the variance estimates are unacceptably large. This is particularly true for the direct method.

Figure 6.10: *Boxplots showing the range of estimates of* $\mathbb{V}_k[\,\hat{\beta}^s\,]$ *and* $\mathbb{V}_e[\,\hat{\beta}^s\,]$*, using both the components method (separately estimating* V_1, V_2, V_3 *and* V_4*) and the direct method (Section 6.4) from 1,000 simulated datasets for hypothetical example (a) of Chapter 5, with a sample size of 2,000.*



6.6 Confidence intervals

We finish this chapter by using these variance estimators to construct confidence intervals for the stratified treatment effect estimator. We compare various normal-based confidence intervals of the form

$$\left(\hat{\beta}^s - 1.96 \times \sqrt{V},\ \hat{\beta}^s + 1.96 \times \sqrt{V}\right),$$

where V is some variance estimator for $\hat{\beta}^s$. Five variance estimators are considered: $\mathbb{V}_k[\,\hat{\beta}^s\,]$ estimated using the components method, $\mathbb{V}_k[\,\hat{\beta}^s\,]$ estimated using the direct method, $\mathbb{V}_e[\,\hat{\beta}^s\,]$ estimated using the components method, $\mathbb{V}_e[\,\hat{\beta}^s\,]$ estimated using the direct method, and a pragmatic approach to the estimation of $\mathbb{V}_e[\,\hat{\beta}^s\,]$ that uses the direct method of estimation but replaces the variance estimator with an estimator of $\mathbb{V}_k[\,\hat{\beta}^s\,]$ when the latter is smaller. For brevity, we refer to the resulting confidence intervals as C_k, D_k, C_e, D_e, P_e, where the C refers to the components method, D refers

Figure 6.11: *Boxplots showing the range of estimates of* $\mathbb{V}_k[\,\hat{\beta}^s\,]$ *and* $\mathbb{V}_e[\,\hat{\beta}^s\,]$*, using both the components method (separately estimating* V_1, V_2, V_3 *and* V_4*) and the direct method (Section 6.4) from 1,000 simulated datasets for hypothetical example (b) of Chapter 5, with a sample size of 2,000.*

to the direct method and P refers to the pragmatic method and as before the k and e refer to the propensity score being assumed to be known or estimated, respectively.

An alternative method for constructing confidence intervals for the stratified treatment effect estimator would be to use bootstrap methods [20]. For comparison, we use the percentile and bias-corrected and accelerated methods in our simulations.

We simulated $1,000$ datasets each of size $n = 2,000$ from hypothetical example (a) of Chapter 5. For each of these datasets, seven 95% confidence intervals were calculated: the percentile and the bias-corrected and accelerated bootstrap confidence intervals, C_k, D_k, C_e, D_e, and P_e. Table 6.1 shows the resulting coverage probabilities, and the average confidence interval length.

We see that all seven methods have good coverage properties. The smallest average confidence interval length is achieved by the C_e confidence interval. This is the confidence interval that uses the estimate of $\mathbb{V}_e[\,\hat{\beta}^s\,]$ obtained by separately estimating

the four variance components. Confidence interval D_e — the method which uses the estimate of $\mathbb{V}_e[\,\hat{\beta}^s\,]$ obtained by the direct method — has large variability. This is due to the problem discussed previously in this chapter with the estimation of the variance using the direct method in some examples. Confidence interval P_e — the method which uses the estimate of $\mathbb{V}_e[\,\hat{\beta}^s\,]$ obtained by the direct method but substitutes the estimate of $\mathbb{V}_k[\,\hat{\beta}^s\,]$ when the latter is smaller — also has good coverage, small average length and small standard deviation of the length.

Table 6.1: *95% confidence intervals for hypothetical example (a) of Chapter 5, using 1,000 simulated datasets of size 2,000.*

Method	Coverage (%)	Average length of C.I. (standard deviation)
Percentile	94.6	2.230 (0.10)
Bias-corrected and accelerated	94.6	2.231 (0.10)
C_k	96.0	2.233 (0.09)
C_e	95.4	2.167 (0.07)
D_k	96.0	2.222 (0.09)
D_e	95.5	2.455 (3.93)
P_e	95.4	2.179 (0.07)

6.7 Discussion

In this chapter we have seen that the variance of the stratified treatment effect estimator can be estimated from a sample dataset using kernel density estimation methods. This estimation process appears to work relatively well, especially given the inherent variability of the sample estimate of the component V_1, which is the variance estimator that is typically used in practice [59]. At present, however, it is not clear how we can tell in advance whether the kernel density estimation will perform well on a particular dataset.

In the simulation study comparing confidence intervals obtained with these variance estimators with bootstrap confidence intervals, both performed well. The average length of the confidence interval was smallest for the method that used a variance estimate obtained by estimating the four variance components separately. Using the estimate of the variance obtained from the direct method but replacing this with an estimate of $\mathbb{V}_k[\,\hat{\beta}^s\,]$ when the latter was smaller seems to be a promising method. The

direct method has the advantage that it could be extended to different situations, such as non-continuous outcomes, fairly easily.

So far we have only used these methods on hypothetical simulated datasets. In the following chapter, therefore, we apply these methods to a real dataset.

Chapter 7

Application to the ESCAPE dataset

We have so far calculated formulæ for the variance of the stratified treatment effect estimator both when the propensity score is known and when it is estimated using a correctly specified logistic regression model. We have shown how kernel density estimation methods can be used to estimate these variances from hypothetical sample datasets. We now apply these methods to a real dataset. The data is taken from a cluster randomised trial — the ESCAPE trial — designed to investigate the effect of an exercise programme on disability, for elderly patients with knee pain.

7.1 Introduction

The high prevalence of chronic knee pain in the elderly [74] leads to substantial disability and socioeconomic costs [60]. As the population ages this problem will increase. There are concerns about the efficacy and side-effects associated with the palliative drugs typically used to treat knee pain [51]. Furthermore, lessening pain will not necessarily decrease disability, which is often as important as pain [45]. There is therefore a need for safe and practical interventions that can lessen pain and disability in elderly patients with chronic knee pain.

It has been shown that psychosocial variables, such as depression and self-efficacy, are associated with perceived levels of pain, mood, and coping efficacy [53]. If these psychosocial variables can be manipulated through, for example, counselling and stress-management classes then it may be possible to reduce pain and disability of elderly patients with chronic knee pain in a safe and effective manner, removing some of the associated socioeconomic costs.

The psychosocial variable we investigate here is the patient's belief in their ability to influence their condition through exercise [28]. This is comprised of four components:

the patient's belief in their ability to exercise (self-efficacy for exercise), the patient's perception of barriers to exercise, the patient's expectations of the benefit of exercise, and the patient's expectations of the impact of exercise. Our primary hypothesis is that the higher a patient's belief in their ability to influence their condition through exercise, the lower their self-reported disability in the future.

In order to investigate this question, data was taken from a cluster randomised trial, which was designed to estimate the effect of a personalised, progressive rehabilitation program on disability, when compared with standard GP care.

7.2 Methods

7.2.1 The ESCAPE dataset

The data is taken from the ESCAPE cluster randomised trial, details of which can be found elsewhere [45]. Data was collected on 418 patients from 54 different surgeries, where the surgeries were block randomised in groups of three to receive one of three interventions: usual GP care, an individual rehabilitation program or a group rehabilitation program. The rehabilitation program, whether administered in groups (approx. 8-10 patients) or individually, consisted of 12 supervised sessions over the course of 6 weeks. During each session, the physiotherapist facilitated a discussion on a topic such as diet or pain control. This was followed by exercises focussing on balance, control and function. The trial recruited people aged 50 years or older who had consulted their primary care physician previously for recurrent knee pain, without excluding those with stable co-morbidities common to that age group. People were excluded if they had had physiotherapy for knee pain in the preceding 12 months, intra-articular injections in the preceding 6 months, lower limb arthroplasty or unstable medical conditions. Also excluded were those who were unable or unwilling to exercise, patients with a severe lack of mobility and patients who were unable to understand English.

Primary outcome. The Western Ontario and McMasters University Osteoarthritis Index (WOMAC) [8] was administered at baseline and approximately six months later. The questionnaire produces a total score consisting of three sub-scores: physical function, pain and stiffness. Our primary outcome is the sub-score measuring physical

function, or disability (WOMAC-function). Higher values on this scale indicate lower function (more disability).

Exposure of interest. A questionnaire asking about the patient's belief in their ability to influence their condition through exercise (Exercise Beliefs) was also administered at baseline [28]. This gave a score comprised of four sub-scores: belief in ability to perform the exercise, perceptions of barriers to exercise, expectations of the benefit of exercise, and expectations of the impact of exercising. The total of these four sub-scores measures the patient's overall belief in their ability to influence their condition through exercise. Higher values on this scale indicate that the patient has more confidence in their ability to improve their knee condition. This is a slightly broader concept than self-efficacy which, strictly speaking, is measured by the first of the four sub-scores mentioned. Following Maliski *et al.* [61] we dichotomize the total score into low and high exercise beliefs on the basis of the empirical sample distribution. Since these exercise beliefs are not randomly allocated and are likely to be associated with other factors that affect the outcome, the ESCAPE dataset is then effectively an observational dataset within which some subjects also had access to an exercise program.

Baseline data was collected on the following variables: sex, age, height, weight, BMI, duration of symptoms, aggregate functional performance time (AFPT) — an objective measure of the time taken to perform certain routine tasks, hospital depression and anxiety score (HAD-depression, HAD-anxiety) [116], and condition-specific health related quality of life (MACTAR) [13].

Patients who had missing data for the outcome, exposure or any of the potential confounders listed were excluded from the following analyses.

Beliefs about exercise may influence function in various ways, one of which is through willingness to take advantage of available methods of exercise. For patients who were allocated to a rehabilitation arm of the trial, the number of exercise sessions attended (which ranged from 0 to 12) is likely to be associated with beliefs about exercise. Therefore, although we adjust for allocation to a rehabilitation arm of the trial, we do not adjust for the number of exercise sessions attended as this is likely to be on the causal pathway.

All confidence intervals calculated in the following analyses are 95% confidence intervals.

7.2.2 Outcome regression analysis

The ESCAPE data is hierarchical since patients are clustered within GP surgeries. However, the original analysis of the dataset [45] found no evidence of a surgery-level effect. We therefore begin by analysing the data as if it were independent. We later perform a sensitivity analysis to assess the effect of ignoring the clustering.

An ordinary least-squares model was fitted to estimate the effect of the dichotomized exercise beliefs on WOMAC-function score at six months, adjusting for a set of covariates. Possible confounders of the relationship between exercise beliefs and function, measured at baseline, are: WOMAC-function, WOMAC-pain, WOMAC-stiffness, HAD-depression, HAD-anxiety, sex, age, height, weight, BMI, duration of symptoms, MACTAR quality of life, and allocation to rehabilitation. The selection of covariates to include in the regression model was done in two ways. Firstly, a model containing all candidate covariates was fitted. Secondly, backward selection was used to select a model from the list of candidate covariates above plus quadratic and logarithmic terms for each continuous covariate and also one-way interactions with age, using a p-value of 0.05 for retention in the model. Standard model checks were carried out for both models, using graphical methods and the Cook-Weisberg test for heteroskedasticity.

7.2.3 Propensity score analysis

The propensity score was estimated using a logistic regression model of dichotomized exercise beliefs on the following baseline covariates: WOMAC-function, WOMAC-pain, WOMAC-stiffness, HAD-depression, HAD-anxiety, sex, age, height, weight, BMI, duration of symptoms, MACTAR quality of life, and allocation to rehabilitation. Since parsimony is not a goal in a propensity score analysis, it is common to include non-linear terms and interactions. However, there must be few enough covariates for the estimated propensity score to be consistent. Since the ESCAPE dataset is relatively small, with only 418 subjects, non-linear terms and interactions were not included in the model.

The common support condition was imposed by deleting subjects with high exercise beliefs who had a propensity score higher than any subject with low exercise beliefs and deleting subjects with low exercise beliefs who had a lower propensity score than any subject with high exercise beliefs. These subjects are ignored in the analysis since they may be intrinsically non-comparable with any subject in the other exposure group. The propensity score was then re-fitted on the smaller dataset.

The dataset was split into 5 strata based on the quintiles of the estimated propensity score. The distribution of the propensity score was examined using both a histogram and a kernel density estimate. Balance of the distribution of the propensity score within strata between subjects with high and low exercise beliefs was assessed graphically using boxplots, and also analytically using two-sample Kolmogorov-Smirnov tests within strata. Balance of the observed covariates within strata was also assessed by within-strata Kolmogorov-Smirnov tests.

The treatment effect within each strata was estimated by taking the difference in mean WOMAC function, measured at six months, between subjects with high and low exercise beliefs, in each strata. The stratified treatment effect estimate was calculated by taking the unweighted mean of the five within-strata treatment effect estimates. Confidence intervals for the estimate were calculated firstly by using a normal-based confidence interval with the variance derived previously in this thesis, and secondly using the bias-corrected and accelerated bootstrap confidence interval.

7.2.4 Continuous exercise beliefs

In the previous analyses, the exposure variable — exercise beliefs — was dichotomized in order to create two exposure groups. The questionnaire used to measure exercise beliefs initially returned a continuous score. Although in this thesis we have focussed mainly on dichotomous exposure groups, several propensity score methods have been developed to analyse situations where the exposure variable is continuous (see Section 2.4). The method we now apply was developed by Hirano and Imbens [39].

We assume that the exposure Z lies in an interval, $Z \in [\, z_0, z_1 \,]$. The propensity score for the continuous exposure, which Hirano and Imbens term the generalised propensity score, is defined as the conditional density of the exposure given the covariates,

$p(Z, \mathbf{X}) = f_{Z|\mathbf{X}}(Z \mid \mathbf{X})$. The generalised propensity score has a similar balancing property to the usual propensity score, $\mathbf{X} \perp 1_{[Z=z]} \mid p(z, \mathbf{X})$.

The potential outcomes for subject i can be viewed as a function of the continuous exposure, $Y_i(z)$, called the unit-level dose-response function. We wish to estimate the expected outcome in the population that we would see if everyone were given level z of the exposure, $\mathbb{E}[Y_i(z)]$. Analogously to our assumption of strongly ignorable treatment assignment (Assumption 2.2), Hirano and Imbens assume that any potential outcome is independent of the exposure conditional on the generalised propensity score, $Y(z) \perp Z \mid p(z, \mathbf{X})$. Using this assumption,

$$\begin{aligned}\mathbb{E}[Y_i(z)] &= \mathbb{E}_{p(z,\mathbf{X})}[\mathbb{E}[Y_i(z) \mid p(z, \mathbf{X})]] \\ &= \mathbb{E}_{p(z,\mathbf{X})}[\mathbb{E}[Y \mid Z = z, p(z, \mathbf{X})]].\end{aligned}$$

Therefore, there are three steps to estimate $\mathbb{E}[Y_i(z)]$. We begin by estimating the generalised propensity score. We then estimate the outcome given exposure level and generalised propensity score. Finally, we estimate the expectation of this over the distribution of the generalised propensity score.

More specifically, Hirano and Imbens assume that the continuous treatment is normally distributed with some unknown constant variance and a mean that depends linearly on a vector of covariates, $Z \sim N(\boldsymbol{\alpha}^T \mathbf{X}, \sigma^2)$. This model is fitted using least-squares to get estimated parameters $\hat{\boldsymbol{\alpha}}$. Then the estimated generalised propensity score is

$$\hat{p}(z, \mathbf{X}) = \frac{1}{\sqrt{2\pi\,\sigma^2}}\, e^{-\frac{(z-\hat{\boldsymbol{\alpha}}^T \mathbf{X})^2}{2\sigma^2}}.$$

We then need to estimate $\mathbb{E}[Y \mid Z = z, p(z, \mathbf{X})]$. Hirano and Imbens estimate this using the model,

$$\mathbb{E}[Y \mid Z = z, p(z, \mathbf{X}) = p] = \gamma_0 + \gamma_1 z + \gamma_2 z^2 + \gamma_3 p + \gamma_4 p^2 + \gamma_5 z p.$$

Estimates $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3, \hat{\gamma}_4, \hat{\gamma}_5)$ can be obtained by fitting a least-squares model. Then to estimate the expectation of the potential outcome at a particular exposure level, z, we take a sample average,

$$\hat{\mathbb{E}}[Y(z)] = \frac{1}{n}\sum_{i=1}^{n} \left\{\hat{\gamma}_0 + \hat{\gamma}_1 z + \hat{\gamma}_2 z^2 + \hat{\gamma}_3 \hat{p}(z, \mathbf{X}_i) + \hat{\gamma}_4 \hat{p}(z, \mathbf{X}_i)^2 + \hat{\gamma}_5 z \hat{p}(z, \mathbf{X}_i)\right\}.$$

For the ESCAPE data the generalised propensity score was estimated using an ordinary least squares model of (continuous) exercise beliefs on the following baseline covariates: WOMAC-function, WOMAC-pain, WOMAC-stiffness, HAD-depression HAD-anxiety, sex, age, height, weight, BMI, duration of symptoms, MACTAR quality of life, and allocation to rehabilitation. Balance of the covariates was assessed as follows. The sample was split into three groups by the tertiles of the continuous exercise beliefs. The generalised propensity score was estimated at the median of the exercise beliefs in the first of the three tertiles. Then the whole sample was split into five strata by the estimated generalised propensity score. Within each of these five strata, we expect the distribution of, for example, age to be approximately the same in the first tertile as in the second and third tertiles combined. This was assessed using a two-sample Kolmogorov-Smirnov test. The expectation of the potential outcomes, $\mathbb{E}\,[\,Y(z)\,]$, was then estimated at ten values of exercise beliefs, equally spread over the observed range of exercise beliefs. In order to compare this method with the dichotomous propensity score method, the expectation of the potential outcome, $\mathbb{E}\,[\,Y(z)\,]$, was estimated using the generalised propensity score method at the median of each of the two groups of dichotomized exercise beliefs and the difference between these two expected potential outcome was compared with the previously obtained estimated of exposure effect.

7.2.5 Mixed effect models

The previous analyses all assumed that the data was independent, whereas in fact the subjects are clustered within GP surgeries. We now perform both an outcome regression analysis and a propensity score analysis taking account of the clustered nature of the data.

For the outcome regression analysis, a random-effects model of WOMAC-function at three months was fitted including the dichotomous exposure and the following baseline covariates: WOMAC-function, WOMAC-pain, WOMAC-stiffness, HAD-depression, HAD-anxiety, sex, age, height, weight, BMI, duration of symptoms, MACTAR quality of life, and allocation to rehabilitation. A random effect at the surgery level was also added. A Breusch and Pagan test for random effects was used to assess the statistical evidence of clustering.

There is no clear consensus about how to handle clustered data in a propensity score analysis. Some studies have estimated the propensity score separately within each cluster [76], some have ignored the clustering [94], and others have included cluster-level random-effects in the propensity score model. In the ESCAPE dataset, there are insufficient observations in many of the GP surgeries to estimate the propensity score separately within each surgery. The previous analysis ignored the clustering. We now adopt the third approach towards clustered data. The propensity score was modelled using a logistic regression model of dichotomized exercise beliefs on the following covariates: WOMAC-function, WOMAC-pain, WOMAC-stiffness, HAD-depression, HAD-anxiety, sex, age, height, weight, BMI, duration of symptoms, MACTAR quality of life, and allocation to rehabilitation. A random-effect was also added at the surgery level. The resulting estimated propensity score was then used to create five strata and the stratified treatment effect estimate was calculated as before. Since the variance formula calculated previously in this thesis is not valid for clustered data, a normal-based confidence interval was derived using a bootstrap estimate of variance.

7.3 Results

7.3.1 Trial characteristics

Data was collected from 54 GP practices, on 418 patients. By 6-months 81 (19%) participants had withdrawn from the ESCAPE trial. Five of these patients agreed to fill in a postal questionnaire leaving available outcome data for 342 (82%) of the patients. A further 37 (8.9%) have data missing on either the exposure variable — exercise beliefs — or at least one of the following baseline covariates: WOMAC-function, WOMAC-pain, WOMAC-stiffness, HAD-depression HAD-anxiety, sex, age, height, weight, BMI, duration of symptoms, and MACTAR quality of life. This leaves complete data on 305 (73.0%) patients.

Since most variables in this dataset are highly skewed, we report medians and ranges rather than means. The level of exercise beliefs, as measured by the questionnaire, had a median (range) of 69 (65 – 85) in the group with high exercise beliefs, and 60 (42 – 64) in the group with low exercise beliefs. The median (range) of WOMAC-function in subjects with high exercise beliefs was 15 (0 – 57) and the median in subjects with low exercise beliefs was 28 (0 – 62).

Table 7.1 shows the baseline covariates. The demographic covariates are all well balanced between subjects with low and high exercise beliefs. Covariates measuring pain and function, however, are less balanced. Aggregate functional performance time (AFPT) appears to be slightly lower in subjects with high exercise beliefs, indicating higher objective levels of function. Similarly, baseline WOMAC-function appears to be lower in subjects with high exercise beliefs, indicating higher perceived levels of function. Depression and anxiety are marginally lower in those subjects with high exercise beliefs and health-related quality of life is slightly higher for those subjects with high exercise beliefs.

Table 7.1: *Baseline data for subjects with high and low exercise beliefs. Continuous variables are reported as median (range).*

Characteristic	Low self-efficacy $N = 196$	High self-efficacy $N = 214$
Demographic variables		
Sex (% female)	70.6	69.9
Age (years)	66 (50, 91)	66 (51, 90)
Height (metres)	1.62 (1.39, 1.89)	1.65 (1.47, 1.97)
Weight (kg)	80 (48, 139)	79 (47, 133)
BMI	30.5 (21.2, 49.8)	28.4 (19.2, 51.3)
Symptom duration (yrs)	6.0 (0.2, 60.0)	5.0 (0.3, 56.0)
Pain and function		
WOMAC-function	32 (0, 65)	21 (0, 55)
WOMAC-pain	8 (0, 20)	6 (0, 17)
WOMAC-stiffness	4 (0, 8)	4 (0, 8)
AFPT	58.5 (28.0, 282.0)	44.6 (24.4, 225.7)
Depression		
HAD-anxiety	7 (0, 21)	5 (0, 17)
HAD-depression	5 (0, 19)	3 (0, 14)
Health-related quality of life		
MACTAR	30 (19, 40)	33 (21, 40)
Trial variables		
Allocated to treatment (%)	67.3	64.8

7.3.2 Outcome regression analysis

The first outcome regression model of WOMAC-function at six months containing all candidate covariates estimated that the effect of high exercise beliefs was

$-2.92\,(-5.38,\ -0.45)$, indicating an improvement in self-reported function. Standard model checks identified significant evidence of heteroskedasticity, ($p = 0.01$, Cook-Weisberg test). The second model, using a stepwise procedure including all candidate covariates plus quadratic and natural logarithm terms for each continuous covariate and all one-way interactions with sex, selected a model containing 31 covariates. The effect of high exercise beliefs on WOMAC-function at six months was estimated by this model as $-3.43\,(-5.93,\ -0.92)$, a slightly greater estimate of effect than the previous regression model. This second, less parsimonious model also showed some evidence of heteroskedasticity, ($p = 0.06$, Cook-Weisberg test). No other violation of the model assumptions was detected for either outcome regression model.

7.3.3 Propensity score analysis

Imposing the common support condition resulted in the deletion of five subjects. Figure 7.1 shows a histogram of the estimated propensity score, along with a kernel density estimate. The density appears to be fairly continuous with a range from 0.1 to 0.93. Figure 7.2 shows boxplots of the distribution of the estimated propensity score within each of the five strata, for subjects with high and low exercise beliefs. The distribution of the propensity score across exposure groups appears to be fairly balanced within strata, although two-sample Kolmogorov-Smirnov tests indicate some evidence of imbalance in the second and fifth strata ($p = 0.09$, $p = 0.05$). Two-sample Kolmogorov-Smirnov tests (results not shown) indicate no evidence of within-strata covariate imbalance across exposure groups among the covariates measuring pain and function, although there is some evidence of imbalance of the duration of symptoms.

Figure 7.1: *Histogram and kernel density estimate of the propensity score.*

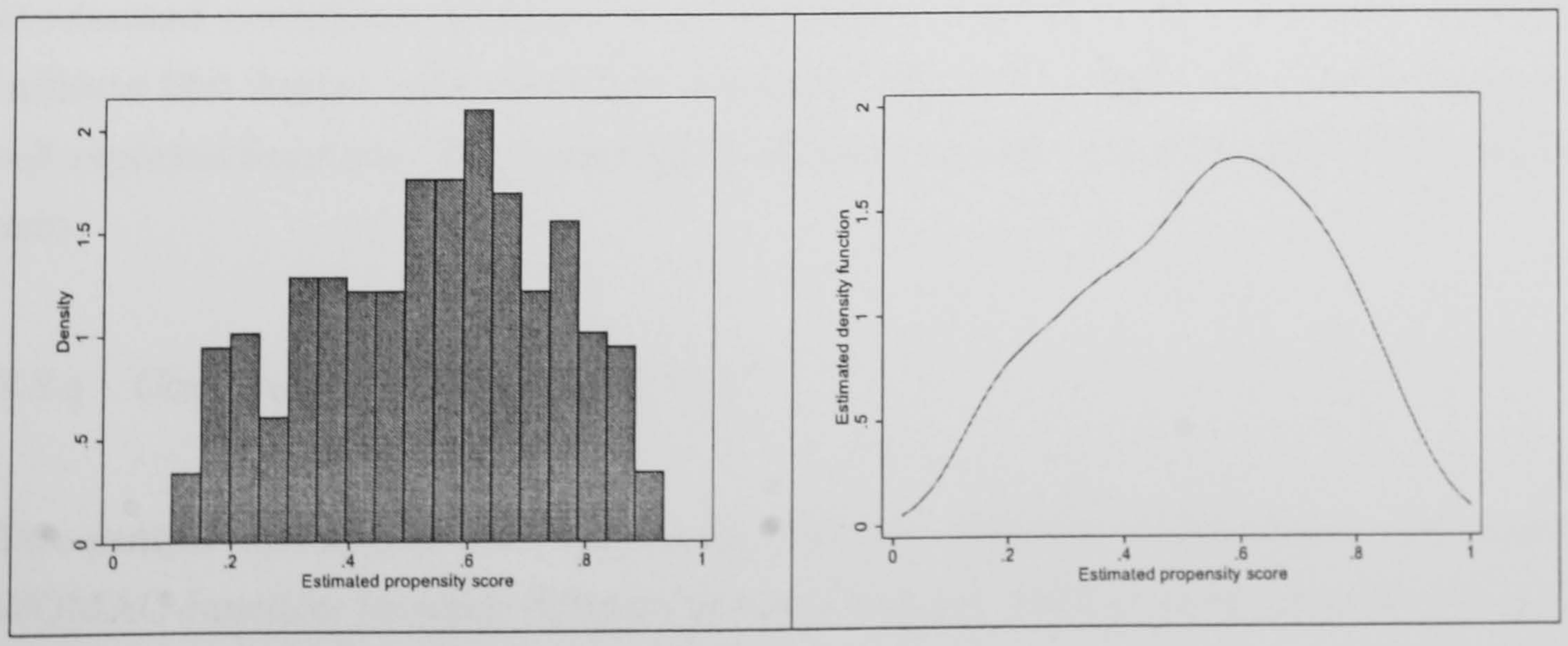
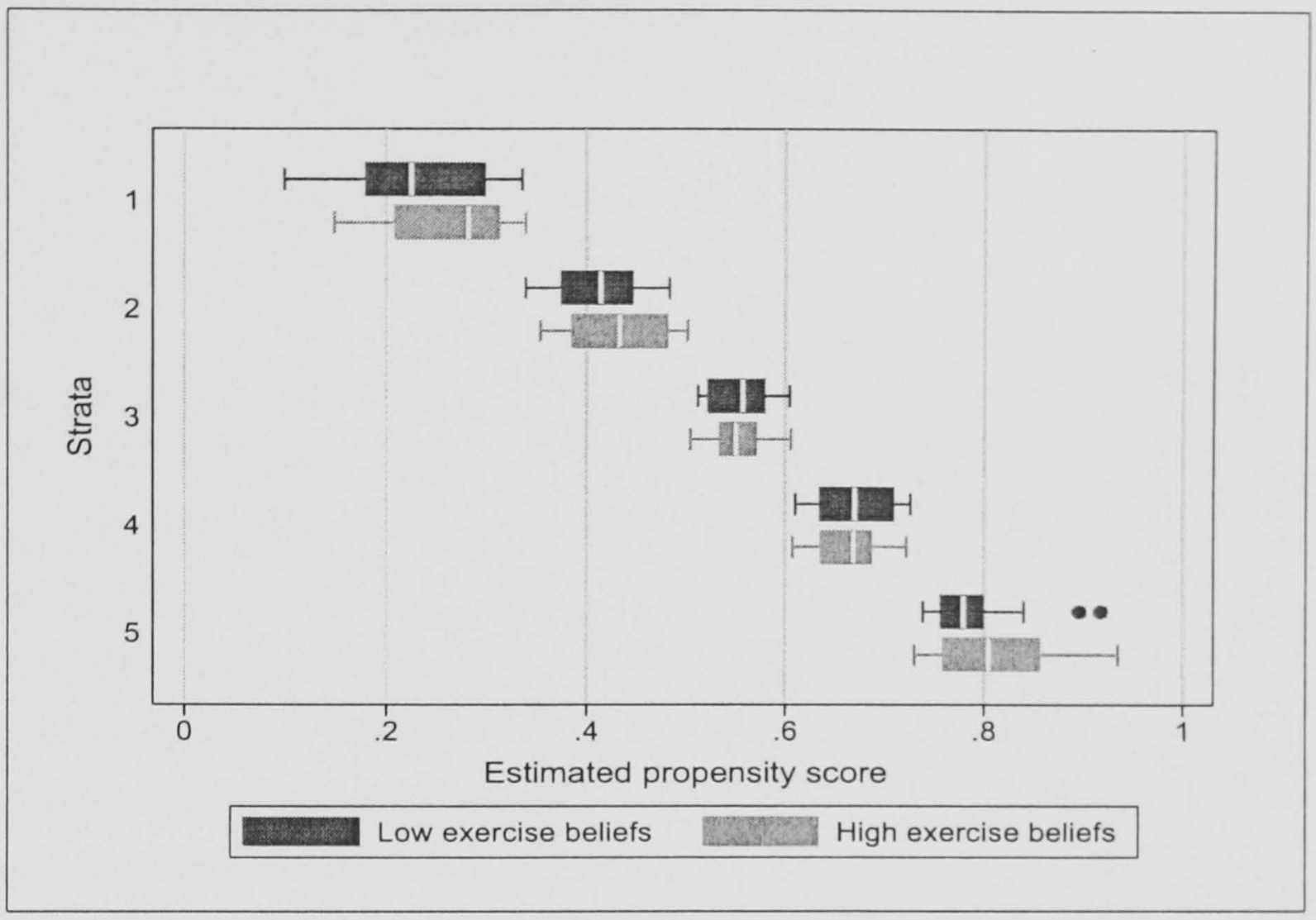


Figure 7.2: *Boxplots of the estimated propensity score within strata, for subjects with high and low exercise beliefs.*



Stratification by this propensity score, using five equal-sized strata, estimated that the effect of high exercise beliefs on WOMAC-function at six months was -2.99. Estimated variance components are shown in Table 7.2. Estimation of the propensity score is estimated to have reduced the variance of the effect estimate by about 34.9%. Normal-based confidence intervals obtained using the variance estimated by the sum of the four variance components and using the variance estimated by the direct method described in Chapter 6 are respectively: $(-5.60, -0.37)$ and $(-5.73, -0.24)$. All conditions required for the validity of the variance formulæ calculated earlier in the thesis (listed in detail in Chapter 3) appeared to be satisfied. The bias-corrected and accelerated confidence interval is: $(-6.07, 0.19)$. The first two confidence intervals indicate that higher exercise beliefs are associated with a significant improvement in self-reported function. The bootstrap confidence interval is slightly wider and includes zero.

7.3.4 Continuous exercise beliefs

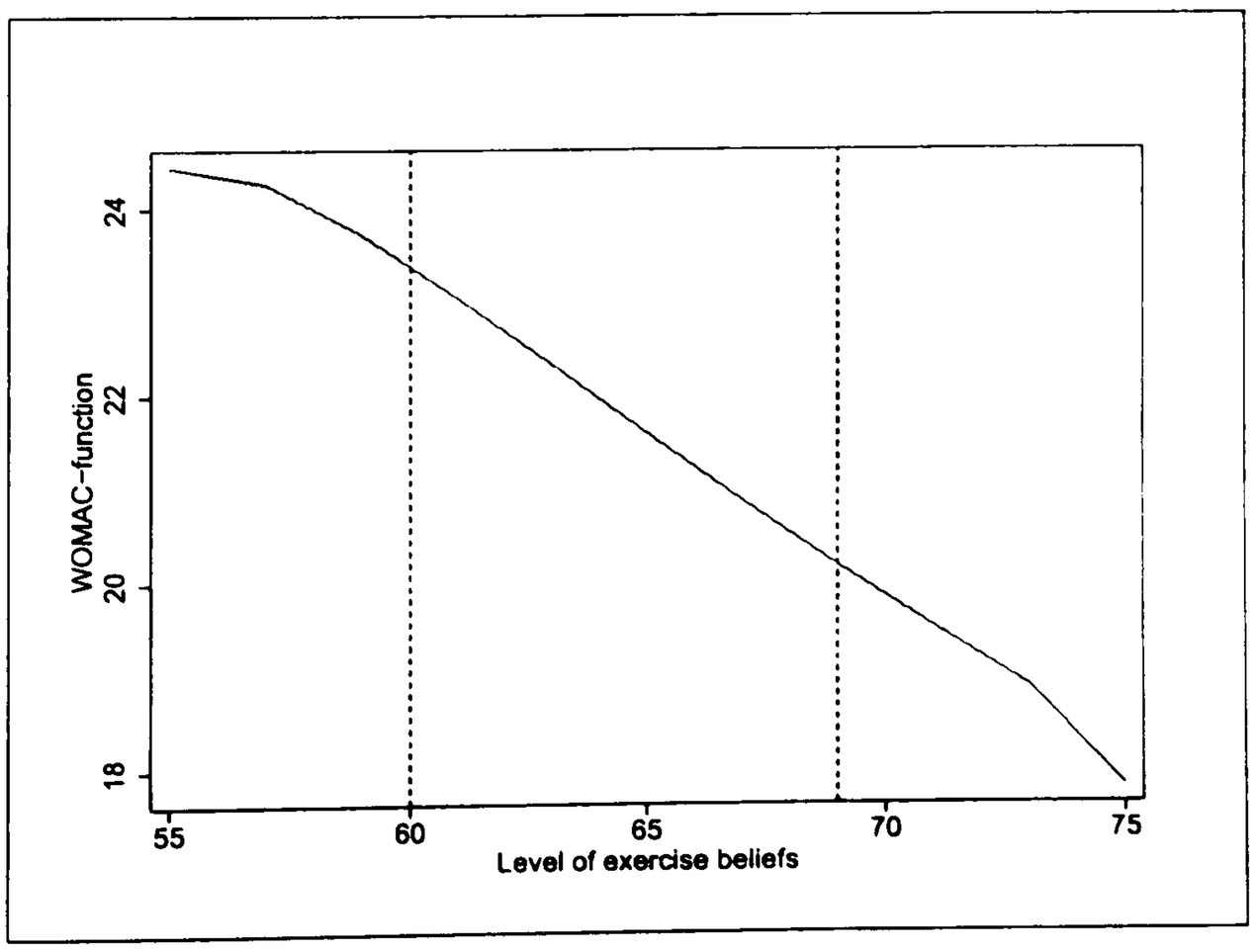
Two-sample Kolmogorov-Smirnov tests indicated evidence of imbalance of baseline WOMAC-function between different levels of exercise beliefs, conditional on the gen-

Table 7.2: *Estimated variance components for the stratified estimate of the effect of high exercise beliefs on WOMAC-function at 6 months.*

Component	Estimate
V_1	815.6
V_2	0.11
V_3	−284.4
V_4	4.03

eralised propensity score, (p = 0.04). Therefore, baseline WOMAC-function was added to the regression model of WOMAC-function at six months on the generalised propensity score and exercise beliefs. This resulted in the estimated dose-response function shown in Figure (7.3). The relationship between exercise beliefs and WOMAC-function appears to be fairly linear, with higher exercise beliefs being associated with lower WOMAC-function (less disability). In order to compare this method with the previous propensity score analysis, the dose-response function was evaluated at levels of exercise beliefs of 60 and 69, the median exercise beliefs in the high and low exercise belief groups used in previous analyses. This estimated that the effect of high versus low exercise beliefs on WOMAC-function at six months was -3.22.

Figure 7.3: *Estimated dose-response function for WOMAC-function at six months for a range of exercise beliefs.*



7.3.5 Mixed-effects models

A mixed-effects model of WOMAC-function at six months including all candidate covariates, the dichotomised exercise beliefs, and a random effect of surgery estimated that the effect of high exercise beliefs was $-2.51\,(-4.96, -0.06)$. The Breusch and Pagan test for random effects showed no evidence of a surgery level effect ($p = 0.76$).

The propensity score was modelled as before but with the addition of a random effect for surgery. The stratified treatment effect estimate was calculated as before with the resulting estimated propensity score. As before, two-sample Kolmogorov-Smirnov tests showed no evidence of imbalance of covariates within strata between subjects with high and low exercise beliefs, other than the duration of symptoms. Since the variance formula calculated earlier in this thesis assumed that the propensity score was calculated using a logistic regression model with no random effects, a normal-based bootstrap confidence interval was calculated. This analysis estimated that the effect of high exercise beliefs on WOMAC-function at six months was $-3.03\,(-6.02, -0.05)$.

7.4 Discussion

7.4.1 Comparison of methods

Table 7.3 shows the estimates of the effect of high exercise beliefs on WOMAC-function at six months with 95% confidence intervals from each of the methods used in this analysis. All methods produce similar point estimates. The highest estimate of effect comes from the non-parsimonious outcome regression model, and the lowest comes from the outcome regression model including a surgery-level random effect. Only the confidence interval using the bias-corrected and accelerated bootstrap confidence interval in the propensity score analysis includes zero.

For this analysis, the conditions for the validity of the formula for the variance of the stratified treatment effect estimate assuming that the propensity score is estimated using a correctly specified logistic regression model appeared to be satisfied. The resulting variance, when calculated using either the component or the direct method, was similar to the variances from the outcome regression models. However, simulations carried out in Chapter 5 suggest that a sample size of about $n = 2,000$ is needed

Table 7.3: *Point estimates and 95% confidence intervals for the effect of high exercise beliefs on WOMAC-function at 6 months.*

Method	
Outcome regression with all covariates	−2.92 (−5.38, −0.15)
Less parsimonious outcome regression	−3.43 (−5.93, −0.92)
Propensity score (components variance estimate)	−2.99 (−6.29, −0.76)
Propensity score (direct variance estimate)	−2.99 (−6.27, −0.78)
Propensity score (BCa bootstrap variance)	−2.99 (−6.07, 0.19)
Generalised propensity score	−3.22 –
Outcome regression with random effects	−2.51 (−4.96, −0.06)
Propensity score with random effects	−3.03 (−6.02, −0.05)

for the asymptotic variance to be close to the finite-sample variance. Therefore, the bootstrap confidence interval may be preferable in this situation.

The main limitation of this study is the sample size. Data was available only on 418 subjects. Propensity score methods are typically used on larger datasets. Balance is more readily achieved in such datasets and, for example, whilst in such a small dataset a non-significant test for within-strata balance may be due to the small within-strata sample sizes, in a larger dataset any imbalance would be more likely to be detected. The propensity score method used for continuous exposure variables could not balance an important prognostic covariate, baseline WOMAC-function, and so the method had to be adapted to produce a reasonable estimate. In a larger sample this problem would be much less likely to occur.

7.4.2 Possible extensions of the analysis

When analysing an observational dataset that has also been used as part of a randomized trial, it is often possible to use the random treatment allocation as an instrumental variable [101]. In this case, however, the exposure of interest has a very weak correlation with allocation to treatment, which precludes an instrumental variables approach ($p = 0.624$, t-test of continuous exercise beliefs by allocation to treatment).

Various methods have been proposed to deal with missing data in the context of propensity scores. Rosenbaum and Rubin suggest a pattern-mixture model based approach [85]. When there are only missing discrete covariates, this method is equivalent to creating a 'missing' category for each covariate with missing values. Alternatively,

when there are only a few missing data patterns, a separate logistic regression model can be fitted for each pattern. In this dataset, however, the missing values are all in continuous covariates so the first method cannot be used. Furthermore, there are too many different missing data patterns to use the second. More complex alternatives have been proposed [19] but these are often difficult to implement and there is, as yet, no available software that performs such analyses on general datasets. General missing data methods such as multiple imputation would be a better approach with this dataset.

7.4.3 Clinical significance

This study offers additional evidence of the importance of a patient's belief in the efficacy of an intervention on the outcome. In particular, we found that the higher a patient's belief in their ability to improve their chronic knee pain through exercise, the lower their self-reported disability when measured six months later. A patient's belief in their ability to influence their condition is not an unchanging characteristic. It can be modified through, for example, educational classes or stress-management classes. Various self-management programmes, that attempt to enhance coping skills for better symptom management, have been studied and found to be effective [33] [35].

The magnitude of the effect of high exercise beliefs on WOMAC-function is small, in clinical terms. It is, however, comparable with the size of the effect of the exercise program that the ESCAPE trial was designed to assess [45]. If safe but effective interventions to raise beliefs about exercise could be produced, then it might be possible to substantially reduce the disability and the socioeconomic costs associated with chronic knee pain in the elderly.

Chapter 8

Discussion

8.1 Summary

In this thesis we began by placing the various propensity score methods in a clear, cohesive framework in order to better understand how they relate to each other. This was used to consider how we might expect each method to behave in different situations.

We then moved on to ascertain the theoretical properties of the stratified treatment effect estimator, the estimator obtained by stratification on the propensity score. We derived conditions under which this estimator is consistent and asymptotically normal. We then calculated its variance, demonstrating that this variance is distinct from the variance used routinely in epidemiological applications. Simulation studies suggest that the variance of the stratified treatment effect estimator, assuming that the propensity score is estimated using a correctly specified logistic regression model, is always smaller than the variance when the propensity score is known. This, in turn, suggests that the variance used in applications is producing hypothesis tests and confidence intervals that are too conservative.

We calculated the conditional variance of the stratified treatment effect estimator given the observed treatment and covariates and, by marginalising this conditional variance, demonstrated that the variances calculated previously in the thesis are asymptotic marginal variances.

We developed a method for estimating the variance of the stratified treatment effect estimator from a sample dataset using kernel density estimation methods, which we called the components method. For most datasets this estimates the variance well. However, occasionally the variance is greatly overestimated. It is not, at present, clear when this is likely to happen or how we can avoid this problem.

We also developed a second, quicker, but less consistent method of estimating the variance of the stratified treatment effect estimator, which we referred to as the direct method. This method produces an unacceptably large variance estimate more frequently than the components method.

Simulation studies were used to compare the frequentist properties of normal-based confidence intervals for the stratified treatment effect estimate using the variance formulæ derived in this thesis and bootstrap confidence intervals. Despite the odd large variance estimate, the confidence intervals constructed using the variance estimated by the components method appeared to have good theoretical properties, as did the pragmatic (but biased) approach of taking the direct estimate of variance but substituting the routinely-used variance when the direct estimate was obviously too large.

We ended by applying these methods to an example dataset, demonstrating that the results derived in this thesis produced similar inference to a well conducted outcome regression analysis.

8.2 Strengths and weaknesses of this thesis

8.2.1 Strengths

The framework proposed in Chapter 2 clearly ties the various propensity score methods together, allowing hypotheses to be more easily formed about the behaviour that can be expected from each of the methods in various situations. This was applied to the issue of the implications of estimating the propensity score. A similar approach can also be used to think about the effect of, for example, non-linearities and interactions on each of the methods.

The explicit variance formula for the stratified treatment effect estimator clearly shows the four sources of variance that effect the estimator. This throws up some unexpected conditions about, for example, the placing of strata boundaries. In particular, we found that when viewed from a marginal frequentist perspective, placing a strata boundary at a point in which the probability distribution of the propensity score is low

often results in a large increase in both the finite-sample and the asymptotic variance. The mathematical discussion of the four variance components clearly shows that the estimation of the strata boundaries will always increase the variance, although we show that the increase can be expected to be negligible. The discussion of the third and fourth variance, along with empirical evidence both in this thesis and in other studies, provides a reasonable argument that the estimation of the propensity score will always reduce the variance of the stratified treatment effect estimate.

The careful calculation of the conditional variance, and its marginalisation over the distribution of treatment and covariates, shows the link, in this problem, between the variance formula used typically in applications of stratification on the propensity score, conditional variance estimates such as those typically used in regression models, and marginal variance estimates such as the ones calculated in this thesis.

8.2.2 Weaknesses

Although we have focussed, for most of this work, on stratification on the propensity score, there is growing interest in the doubly robust methods of analysis. It is likely that this method produces less biased and less variable estimates than does stratification on the propensity score. However, despite this limitation of the analysis method studied, whilst epidemiologists continue to use stratification on the propensity score it is important to develop methods that improve the inferences obtained using that method as well as developing new methods.

The marginal variance of the stratified treatment effect estimate was calculated in this thesis. Some epidemiologists would argue that we should only attempt to estimate causal parameters for the sample at hand [69] and therefore we would use a conditional variance, given the observed covariate distributions. However, we often wish to generalise our results to a wider population. For example, in public health, we may wish to estimate the treatment effect we would see if we chose to make an intervention available throughout the country, in which case the most appropriate variance would be the marginal variance. Moreover, conditioning on a variable that has no intrinsic meaning seems rather arbitrary, whereas marginalising over it is philosophically a more attractive approach.

One of the initial aims of this thesis was to use the variance formula to show that the estimation of the propensity score always reduces the variance of the stratified treatment effect estimator — a result which has been proven for other propensity score methods. Unfortunately, this was not possible due to the complex nature of the variance formula.

The variance formula derived in this thesis, when the propensity score is estimated, is rather complex and involves unknown derivatives. The methods developed to estimate the variance using kernel density estimation are promising but still have several drawbacks. The variance is occasionally drastically overestimated, especially with the second method of estimation, the direct method. It is not clear at present whether we can reduce this error in any way or whether we could develop methods that could tell us whether the kernel density estimation methods are likely to work for a particular dataset. A bootstrap estimate of variance, or confidence interval, has attractive theoretical properties in this situation and may be an easier and more practical solution.

The simulation studies used to assess the accuracy and the convergence rates of the variance formulæ all contained the same structure of covariates so these methods have not been tested on a wide variety of situations. Furthermore, in all examples used, the propensity scores were created from a combination of normal covariates. These examples are therefore likely to show the gaussian-kernel-based methods at their best. It would be sensible to try the methods on propensity score distributions that are less close to the normal distribution.

The variance formulæ calculated in this thesis are not easily extendible to more complex situations. The direct method is more promising in this respect, but there is still work to be done making this method less prone to substantial error if this is to be applied to other situations.

8.3 Further work

Further work is needed looking at the kernel density estimation method to estimate the variance of the stratified treatment effect estimate. For the components method, a promising idea is to use kernel density estimation techniques to estimate the third variance component as well as the fourth. Since these two components overlap greatly

in terms of the error measured, the two terms substantially cancel one another out. If both components are estimated with error where the error is in the same direction for each, then the subtraction of one estimate from the other will cancel out some of the error. Simulation studies are needed to assess the large sample properties of such a method. This solution will not increase the accuracy of the direct method. Methods that can identify situations where the direct method is likely to fail or that reduce the magnitude or frequency of the mis-estimation of the variance using the direct method would be useful.

The variance formulæ derived in this thesis can be applied to the estimation of $\mathbb{E}[Y_1] - \mathbb{E}[Y_0]$ for any type of data. If the data is continuous, this is the usual estimate of treatment effect. If the data is binary, this is the risk difference. Whilst the risk difference can only be used in observational datasets that are representative samples from the population, when this is the case, the risk difference may be the quantity of interest, in which case the variance formulæ derived in this thesis can be used. Note, however, that with small samples or probabilities of success (or failure) close to zero or one, the normal-based confidence intervals may cross zero or one.

The variance calculation in this thesis could be extended to calculate the marginal variance for the estimate of an odds ratio, using stratification on the propensity score. However, given the results of this calculation, it may be more sensible to use bootstrap methods, since it is likely that the explicit variance formula for the odds ratio variance will also be complex and hard to estimate.

A Bayesian approach would be possible here, which would avoid the problem of the estimation of the variance components since we would estimate the variance by the variance of the posterior distribution. However, the goal here is to provide a simple method of analysis that is clear and easily applied. Whilst the kernel density estimation of the variance components is not simple, the frequentist point estimate is simple, and an epidemiologist using the method would not need to understand the kernel density estimation methods to make inferences from the resulting confidence intervals. A Bayesian analysis is more complex to apply and often involves specialist software. Therefore, we do not pursue the Bayesian option further.

On a more general level, more work is needed to look at situations in which propensity score methods are likely to produce 'better' results than the standard regression

analyses. Also, studies comparing the various propensity score methods, looking at situations where one method outperforms the others, would be instructive. In particular, the discussion in Chapter 2 suggests that the covariate-adjustment method of including the propensity score as a covariate is both the most likely to produce biased estimates of effect and the most popular choice of propensity score method. Clear guidelines indicating when each propensity score method is likely to work well would be very beneficial for epidemiologists and statisticians wishing to use propensity score methodology to estimate causal effects.

8.4 Practical implications for epidemiologists

We would recommend that propensity score methods be used when the treatment or exposure is common, and thus the propensity score can be well estimated, and the assumptions of a standard outcome regression model may be questionable. For example, when there are non-linear relationships between covariates and the outcome, it may be easier to use a propensity score than to correctly model the non-linearity. Situations in which the outcome is rare but the treatment is common are also suitable for propensity score analyses.

When using the method of stratification on the propensity score, we would recommend that the variance $\mathbb{V}_e[\,\hat{\beta}^s\,]$ should be used in applications, rather than $\mathbb{V}_k[\,\hat{\beta}^s\,]$ as is routinely used at the moment, since the latter appears to produce conservative hypothesis tests and confidence intervals. It is fairly easy to violate the conditions necessary for the validity of the variance formulæ, $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$. In most cases, however, we would expect these violations to be apparent during the analysis of the dataset.

The investigation of the convergence rates of the variance formulæ suggest that, ideally, a sample size of at least 2,000 should be used for the variance estimator to be valid. Therefore, with small sample sizes we recommend that a bootstrap estimate of variance should be used. However, in our example dataset, the variance estimator of $\mathbb{V}_e[\,\hat{\beta}^s\,]$ was very similar to that obtained using standard regression analysis. Since the modelling assumptions appeared to be fairly well satisfied in this case, this suggests that our variance estimator was valid in this dataset, even with such a small sample size.

The confidence intervals using kernel density estimation are appealing in that they are non-parametric, much quicker than bootstrap confidence intervals, and there is some indication that they have nice theoretical properties — nominal coverage and a smaller average confidence interval length than the bootstrap confidence intervals. In particular, this will be useful in simulation studies comparing stratification on the propensity score with other methods since, previously, such simulations were limited by computation time. When the conditions for the validity of the variance formula $\mathbb{V}_e[\hat{\beta}^s]$ appear to be satisfied, and the sample size is relatively large, we would recommend use of the components method to derive a confidence interval for the stratified treatment effect estimate. When the conditions are not satisfied, and the problem cannot be modified to satisfy them, bootstrap methods can be used. This, however, is very computer-intensive when large datasets are used. In such situations, a jack-knife estimate of variance is likely to be the best approach.

If using a marginal frequentist approach to inference, it is important to make sure that the strata boundaries do not fall in areas where the probability density function of the propensity score is particularly low, as this results in large increases in variance.

Although various missing data methods have been developed to use within a propensity score context, we found that in practice their application is limited. Until software is available to help epidemiologists and statisticians implement these methods, it seems at present easier to use standard missing data methods, such as multiple imputation.

Appendix A

Proof of Theorem 3.1

A.1 Introduction

In this appendix we investigate the asymptotic properties of the stratified treatment effect estimator, $\hat{\beta}^s$, assuming that the propensity score is a known function of the observed covariates. We begin, in Section A.2, by using standard asymptotic theory to demonstrate that $\hat{\beta}^s$ is a consistent estimator of β_o^s [1]. We then, in Section A.3, show that $\hat{\beta}^s$ is asymptotically normally distributed. We finish by determining its asymptotic variance using M-estimation theory, in Section A.4.

This appendix uses the notation given in Section 3.1.1. In order to ease our way into the proof, we give a brief review of the estimator, $\hat{\beta}^s$, and the estimating equations used to obtain it.

A.1.1 Estimating the stratified treatment effect

In the notation of Section 3.1.1, where, for subject i, Y_i and Z_i denote the outcome and treatment, $\hat{S_{si}} = 1_{[\hat{q}_{(s-1)} \leq p_o(\mathbf{X}) < \hat{q}_s]}$ is an indicator for the s^{th} sample stratum, r_s is the fraction of the sample in the s^{th} stratum, and $\hat{d}_s$ is the fraction of the sample who are treated and in the s^{th} sample stratum, the stratified treatment effect estimator is

$$\hat{\beta}^s = \frac{1}{n} \sum_{s=1}^{K} r_s \sum_{i=1}^{n} \left(\frac{Y_i Z_i \hat{S}_{si}}{\hat{d}_s} - \frac{Y_i (1 - Z_i) \hat{S}_{si}}{r_s - \hat{d}_s} \right).$$

[1] We show that $\hat{\beta}^s$ is consistent for the population parameter β_o^s. Since this is usually different from the population average causal treatment effect, β_o, this means that $\hat{\beta}^s$ will usually not be consistent for β_o.

As discussed in Chapter 3, we obtain $\hat{\beta}^s$ as a component of the vector solution to the set of estimating equations

$$\sum_{i=1}^{n} \psi(Y_i, Z_i, \mathbf{X}_i\,; \beta^s, \mathbf{d}, \mathbf{q}) = 0 \tag{A.1}$$

We call the solution $\hat{\boldsymbol{\theta}}$, defined by $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T)$, where the population values of these parameters are: $\mathbf{q}_o$, the population strata boundaries, $\mathbf{d}_o$, the population probabilities of being both treated and in each stratum, and β^s_o, the 'true' stratified treatment effect.

The estimating equations (A.1) are defined by $\psi^{2K\times 1}$ where $\psi^T = (\psi_1, \psi_2^T, \psi_3^T)$, with

$$\psi_1^{1\times 1}\,(Y_i, Z_i, \mathbf{X}_i\,; \beta^s, \mathbf{d}, \mathbf{q}) = \left(\textstyle\sum_{s=1}^{K} r_s \left\{ \frac{Y_i Z_i S_{si}}{d_s} - \frac{Y_i(1-Z_i)S_{si}}{r_s - d_s} \right\} - \beta^s \right)$$

$$\psi_2^{K\times 1}\,(Z_i, \mathbf{X}_i\,; \mathbf{d}, \mathbf{q}) = \begin{pmatrix} Z_i\, S_{1i} - d_1 \\ \vdots \\ Z_i\, S_{Ki} - d_K \end{pmatrix}$$

$$\psi_3^{(K-1)\times 1}\,(\mathbf{X}_i\,; \mathbf{q}) = \begin{pmatrix} S_{1i} - r_1 \\ \vdots \\ S_{(K-1)i} - r_{(K-1)} \end{pmatrix},$$

where $S_{si} = 1_{[\,q_{(s-1)} \le p_o(\mathbf{X}) < q_s\,]}$ is an indicator for the s^{th} stratum. We define $q_0 = 0$ and $q_K = 1$. We now use this representation of the estimation process to ascertain the theoretical properties of the stratified treatment effect estimator, $\hat{\beta}^s$.

A.2 Consistency

We now show that each component of the estimator $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T)$ is consistent. We begin by establishing the consistency of the estimated strata boundaries, $\hat{\mathbf{q}}$. We then consider the consistency of the estimates of the probabilities of being treated and in each stratum, $\hat{\mathbf{d}}$, treating the strata boundaries as nuisance parameters. Given the consistency of both $\hat{\mathbf{q}}$ and $\hat{\mathbf{d}}$, we then demonstrate the consistency of the stratified treatment effect estimator, $\hat{\beta}^s$.

A.2.1 Consistency of the estimated strata boundaries

Intuitively, we would expect the estimated strata boundaries, $\hat{\mathbf{q}} = (\hat{q}_1, ..., \hat{q}_{K-1})^T$ to be consistent estimators of the population strata boundaries, $\mathbf{q}_o = (q_{1o}, \ldots . q_{(K-1)o})^T$, which are defined as the quantiles of the population distribution of the propensity score that split the population into strata containing fractions $(r_1, ..., r_K)^T$ of the population. We obtain $\hat{\mathbf{q}}$ by solving the estimating equations $\sum_{i=1}^{n} \boldsymbol{\psi}_3(\mathbf{X}_i; \mathbf{q}) = 0$, for $\mathbf{q}$, where $\boldsymbol{\psi}_3$ is defined in Section A.1. We now show that the estimate of the first strata boundary, $\hat{q}_1$, is a consistent estimator of the r_1^{th} quantile of the population distribution of the propensity score, q_{1o}. The consistency of the estimators of all other strata boundaries can be demonstrated in the same way.

The quantile q_{1o} is estimated by solving the first component of the set of estimating equations $\sum_{i=1}^{n} \boldsymbol{\psi}_3(\mathbf{X}_i; \mathbf{q}) = 0$, which is

$$\sum_{i=1}^{n} (S_{1i} - r_1) = \sum_{i=1}^{n} \left(1_{[\,0 \le p_o(\mathbf{X}_i) < q_1\,]} - r_1\right) = 0.$$

Equivalently, if we define

$$\text{sign}\,(a) = \begin{cases} 1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \\ 0 & \text{otherwise} \end{cases}$$

then we can estimate q_{1o} by solving the estimating equation

$$\Psi_n(q_1) \equiv \frac{1}{n} \sum_{i=1}^{n} \text{sign}\,(p_o(\mathbf{X}_i) - q_1) + (2r_1 - 1) = 0.$$

In order to establish consistency of this sample quintile, we now use the following lemma from van der Vaart [108, p.47]. This lemma is concerned with the situation where the estimated parameter, $\hat{\theta}$, is obtained as the solution to an estimating equation $\Psi_n = 0$ where $\Psi_n = \frac{1}{n} \sum_{i=1}^{n} \psi(\theta)$.

Lemma A.1 *Let Θ be a subset of the real line and let Ψ_n be random functions and Ψ a fixed function of θ such that $\Psi_n \xrightarrow{p} \Psi$ for every θ. Assume that each map $\theta \to \Psi_n(\theta)$ is continuous and has exactly one zero, $\hat{\theta}$, or is non-decreasing with $\Psi_n(\hat{\theta}) = o_p(1)$. Let θ_o be a point such that $\Psi(\theta_o - \epsilon) < 0 < \Psi(\theta_o + \epsilon)$ for every $\epsilon > 0$. Then $\hat{\theta} \xrightarrow{p} \theta_o$.*

◇

By the law of large numbers, as $n \to \infty$, for any q_1,

$$\begin{aligned}\Psi_n(q_1) \quad \xrightarrow{p} \quad \Psi(q_1) &= \mathbb{E}\left[\,\text{sign}\,(p_o(\mathbf{X}) - q_1)\,\right] + 2r_1 - 1 \\ &= 2(r_1 - \mathbb{P}(p_o(\mathbf{X}) < q_1)).\end{aligned}$$

The map $q_1 \to \Psi_n(q_1)$ is non-increasing. So in order to prove consistency of $\hat{q}_1$, we merely need to show that there exists a value q_{1o} such that, for every $\epsilon > 0$,

$$\Psi(q_{1o} + \epsilon) < 0 < \Psi(q_{1o} - \epsilon).$$

The equation $\Psi(q_1) = 0$ is solved by q_{1o}, the r_1^{th} quantile of the population distribution of the propensity score, since $\mathbb{P}(p_o(\mathbf{X}) < q_{1o}) = r_1$. If the cumulative distribution function of the propensity score is strictly monotone and continuous at q_{1o} then for every $\epsilon > 0$,

$$\Psi(q_{1o} - \epsilon) = 2(r_1 - \mathbb{P}\left(\,p_o(\mathbf{X}) < q_{1o} - \epsilon\,\right)) > \Psi(q_{1o}).$$

In this way we see that

$$\Psi(q_{1o} + \epsilon) < \quad \Psi(q_{1o}) = 0 \quad < \Psi(q_{1o} - \epsilon).$$

Therefore, applying Lemma A.1 shows that $\hat{q}_1 \xrightarrow{p} q_{1o}$. Under the same condition on the cumulative density function of the propensity score, we note further that the quantile q_{1o} is unique and therefore globally identifiable. This condition will be needed later to demonstrate the consistency of other parameters.

We finally note that the r_s^{th} quantile of the sample distribution of the propensity score, $\hat{q}_s$, can be obtained by solving the estimating equation

$$\Psi_n(q_s) \equiv \sum_{i=1}^{n} \text{sign}\,(p_o(\mathbf{X}_i) - q_s) + (2(r_1 + \ldots + r_s) - 1) = 0.$$

This parameterization removes the apparent dependency of the estimated strata boundaries on the other estimates and so the argument above can again be applied to show that $\hat{q}_s$ is consistent and that q_{so} is globally identifiable.

A.2.2 Consistency of the estimated probabilities of being treated and in each stratum

We now consider the consistency of $\hat{\mathbf{d}} = (\hat{d}_1, .., \hat{d}_K)^T$, the estimated probabilities of being treated and in each stratum. These are obtained by solving the estimating equation $\sum_{i=1}^{n} \psi_2(Z_i, \mathbf{X}_i; \hat{\mathbf{q}}, \mathbf{d}) = 0$ for $\mathbf{d}$, where ψ_2 is defined in Section A.1. The parameters $\hat{\mathbf{d}}$ depend on the estimated strata boundaries, $\hat{\mathbf{q}}$. Therefore, we need to take the estimation of the strata boundaries into account when we consider the consistency of $\hat{\mathbf{d}}$.

Consistency in the presence of a nuisance parameter

In order to demonstrate consistency in the presence of nuisance parameters, we use a theorem from Giurcanu and Trinidade [29]. The theorem considers consistency of an estimator, $\hat{\theta}_2$, under consistent estimation of a nuisance parameter, θ_1. The estimator of interest, $\hat{\theta}_2$, is obtained by finding the value of θ_2 that maximises a (criteria) function $M_n(\hat{\theta}_1, \theta_2)$. The notation $b^* = \arg\max_b M_n(a, b)$ defines b^* as the value of b that maximises the function $M_n(a, b)$.

Theorem A.1 *Let $\boldsymbol{\theta}_o = (\theta_{1o}, \theta_{2o})$ define the population values of the two parameters, assumed to be an interior point of the parameter space, Θ. Let also $M(\theta_1, \theta_2)$ and $M_n(\theta_1, \theta_2)$ be the population and sample criteria functions respectively, where we have $\theta_{2o} = arg\ max_{\theta_2}\ M(\theta_{1o}, \theta_2)$. Suppose that $\hat{\theta}_1$ is a consistent estimator of θ_{1o}, and $\hat{\theta}_2^*(\theta_1) = arg\ max_{\theta_2}\ M_n(\theta_1, \theta_2)$. Assume further that the following conditions hold:*

(i) For all θ_1, $M_n(\theta_1, \theta_2)$ is concave in θ_2 with probability tending to 1;

(ii) $\boldsymbol{\theta}_o = (\theta_{1o}, \theta_{2o})$ is globally identifiable;

(iii) $M(\theta_1, \theta_2)$ is locally Lipschitz in θ_1, uniformly in θ_2, in a neighbourhood of $\boldsymbol{\theta}_o$. That is, for all (θ_1, θ_2) and (θ_1', θ_2) in a neighbourhood of $\boldsymbol{\theta}_o$, there exists a constant k such that $|\, M(\theta_1, \theta_2) - M(\theta_1', \theta_2) \,| \leq k\, |\, \theta_1 - \theta_1' \,|$.

Then $\hat{\theta}_2 \equiv \hat{\theta}_2^*(\hat{\theta}_1) \xrightarrow{p} \theta_{2o}$.

◇

Consistency in the presence of many nuisance parameters

Theorem A.1 deals with the situation where there is a single nuisance parameter. In our problem, however, there are many nuisance parameters and we therefore need to

generalise the theorem to cover these situations. Inspection of the proof of Theorem A.1 shows that the nuisance parameter θ_1 enters the proof only through the condition that for every $\epsilon > 0$, as $n \to \infty$,

$$\mathbb{P}\left(\sup_{\theta_2 \in C} | M(\hat{\theta}_1, \theta_2) - M(\theta_{1o}, \theta_2) | \geq \epsilon/2 \right) \to 0,$$

where C represents a neighbourhood of θ_{2o}. This is ensured by condition *(iii)* of the theorem. We now consider the situation where there are two nuisance parameters, $\boldsymbol{\theta}_1 = (\theta_{1a}, \theta_{1b})$ with population values $\boldsymbol{\theta}_{1o} = (\theta_{1ao}, \theta_{1bo})$, where θ_2 is estimated by maximising the criteria function $M(\theta_{1a}, \theta_{1b}, \theta_2)$. The same proof would show that $\hat{\theta}_2 \equiv \hat{\theta}_2^* (\hat{\theta}_{1a}, \hat{\theta}_{1b}) \xrightarrow{p} \theta_{2o}$, provided that, for every $\epsilon > 0$ as $n \to \infty$,

$$\mathbb{P}\left(S \geq \epsilon/2 \right) \to 0, \tag{A.2}$$

where

$$S = \sup_{\theta_2 \in C} | M(\hat{\theta}_{1a}, \hat{\theta}_{1b}, \theta_2) - M(\theta_{1ao}, \theta_{1bo}, \theta_2) |.$$

Under two Lipschitz conditions, rather than the single one used in Theorem A.1, we can show that (A.2) holds. To begin with,

$$\mathbb{P}\left(S \geq \epsilon/2 \right) \leq \mathbb{P}(S_1 + S_2 \geq \epsilon/2),$$

where

$$S_1 = \sup_{\theta_2 \in C} | M(\hat{\theta}_{1a}, \hat{\theta}_{1b}, \theta_2) - M(\hat{\theta}_{1a}, \theta_{1bo}, \theta_2) |$$
$$S_2 = \sup_{\theta_2 \in C} | M(\hat{\theta}_{1a}, \theta_{1bo}, \theta_2) - M(\theta_{1ao}, \theta_{1bo}, \theta_2) |.$$

Suppose there exist Lipschitz constants k_1 and k_2 such that for all $(\theta_{1a}, \theta_{1b}, \theta_2)$, $(\theta'_{1a}, \theta_{1b}, \theta_2)$ and $(\theta_{1a}, \theta'_{1b}, \theta_2)$ in a neighbourhood of $\boldsymbol{\theta}_o$,

$$| M(\theta_{1a}, \theta_{1b}, \theta_2) - M(\theta'_{1a}, \theta_{1b}, \theta_2) | \leq k_1 | \theta_{1a} - \theta'_{1a} |$$
$$| M(\theta_{1a}, \theta_{1b}, \theta_2) - M(\theta_{1a}, \theta'_{1b}, \theta_2) | \leq k_2 | \theta_{1b} - \theta'_{1b} |.$$

Using the Lipschitz constants k_1 and k_2,

$$\mathbb{P}(S_1 + S_2 \geq \epsilon/2) \leq \mathbb{P}(k_1 | \hat{\theta}_{1a} - \theta_{1ao} | + k_2 | \hat{\theta}_{1b} - \theta_{1bo} | \geq \epsilon/2) \to 0,$$

since $\hat{\theta}_{1b} \to \theta_{1ao}$ and $\hat{\theta}_{1b} \to \theta_{1o}$. Then condition (A.2) is satisfied and so we have the required consistency, $\hat{\theta}_2 \equiv \hat{\theta}_2^* (\hat{\theta}_{1a}, \hat{\theta}_{1b}) \xrightarrow{p} \theta_{2o}$.

Therefore, if we have more than one nuisance parameter, we merely replace condition *(iii)* in Theorem A.1 by the following condition:

(iii) Suppose the nuisance parameter is* $\boldsymbol{\theta}_1^{L\times 1} = (\theta_{11}, ..., \theta_{1L})$ *and the parameter of interest,* θ_2*, is obtained by maximising the sample criteria function* $M_n(\hat{\boldsymbol{\theta}}_1, \theta_2)$*. Then suppose there exists Lipschitz constants* $k_1, ..., k_L$ *such that, for* $l = 1, ..., L$*, for all* $\boldsymbol{\theta}$ *and* $\boldsymbol{\theta}'$ *in a neighbourhood of* $\boldsymbol{\theta}_o$*,*

$$| M(\boldsymbol{\theta}) - M(\boldsymbol{\theta}') | \leq k_l | \theta_{1l} - \theta'_{1l} |,$$

where $\boldsymbol{\theta} = (\theta_{11}, ..., \theta_{1l}, ..., \theta_{1L}, \theta_2)$ and $\boldsymbol{\theta}' = (\theta_{11}, ..., \theta'_{1l}, ..., \theta_{1L}, \theta_2)$. Then provided that all other conditions of Theorem A.1 are satisfied, we have the required consistency, $\hat{\theta}_2 \equiv \hat{\theta}_2^*(\hat{\boldsymbol{\theta}}_1) \xrightarrow{p} \theta_{2o}$.

An application of Theorem A.1

We now apply the generalised version of Theorem A.1 to $\hat{\mathbf{d}} = (\hat{d}_1, ..., \hat{d}_K)^T$, the estimated probabilities of being treated and in each stratum. These probabilities are obtained by solving the set of estimating equations

$$\sum_{i=1}^{n} \boldsymbol{\psi}_2(Z_i, \mathbf{X}_i; \mathbf{d}, \hat{\mathbf{q}}) = 0,$$

where $\boldsymbol{\psi}_2$ is defined in Section A.1. We now demonstrate the consistency of $\hat{d}_1$, obtained by solving the first component of the above estimating equations, which is

$$\sum_{i=1}^{n} (Z_i \ \hat{S}_{1i} - d_1) = 0. \tag{A.3}$$

As before, $\hat{S}_{1i} = 1_{[0 \leq p_o(\mathbf{X}) < \hat{q}_1]}$ is an indicator for the first sample stratum, which is a function of the previously estimated strata boundaries. In order to apply Theorem A.1 to this problem we need to write this estimating equation as a criteria function — a function that is maximised by the same estimate, d_1. Solving (A.3) is equivalent to finding the value of d_1 that maximizes the criteria function

$$M_n(d_1; \hat{q}_1) \equiv -\frac{1}{n} \sum_{i=1}^{n} (\ Z_i \ \hat{S}_{1i} - d_1)^2.$$

We now merely need to verify that all the conditions of Theorem A.1 are satisfied. Firstly, the population parameters d_{1o} and $\mathbf{q}_o$ must be on the interior of the parameter space. Therefore, d_{1o} cannot be equal to either 0 or r_1, and the first population strata boundary cannot be zero or one — a condition which is satisfied since we have already assumed that the propensity score is neither zero nor one for each subject. The function $M_n(d_1; q_1)$ is concave in d_1 and so condition *(i)* of the theorem is satisfied. We have seen that the population strata boundaries are globally identifiable. In order to satisfy condition *(ii)* of the theorem we now show that d_{1o} is globally identifiable. By the law of large numbers, as $n \to \infty$, for any q_1,

$$M_n(d_1; q_1) \xrightarrow{p} M(d_1; q_1) = -\mathbb{E}\,[\,(\,Z\;S_1 - d_1)^2\,].$$

By expanding the expectation on the right-hand side of this equation, we can write

$$M(d_1; q_1) = (1 - 2d_1)\mathbb{E}\,[\,Z\;S_1\,] + d_1^2, \tag{A.4}$$

which can be expressed as

$$M(d_1; q_1) = (d_1 - \mathbb{E}\,[\,Z\;S_1\,])^2 - \mathbb{E}\,[\,Z\;S_1\,](\mathbb{E}\,[\,Z\;S_1\,] - 1).$$

Then $M(d_1; q_{1o})$ is maximised when $d_1 = \mathbb{E}\,[\,Z\;S_{1o}\,]$. This is exactly the definition of the population probability of being treated and in the first population stratum, d_{1o}, which is the parameter we wish to estimate. Therefore, the function $M(d_1; q_{1o})$ is maximised when $d_1 = d_{1o}$. Since this point is unique, d_{1o} is globally identifiable and condition *(ii)* of Theorem A.1 is satisfied.

We have now verified each condition of Theorem A.1 except condition *(iii*)*. In order to verify this condition we need to show that, given $\delta > 0$, for all d_1 and $q_1 \in [\,q_{1o} - \delta, q_{1o} + \delta\,]$, there exists a constant k such that

$$|\,M(d_1; q_1) - M(d_1; q_1')\,| \leq k\,|\,q_1 - q_1'\,|. \tag{A.5}$$

Using (A.4) and conditioning on the observed covariates, $\mathbf{X}$, we have

$$|\,M(d_1; q_1) - M(d_1; q_1')\,| = |\,(1 - 2d_1)\,\mathbb{E}\,[\,p_o(\mathbf{X})(1_{[\,0 \leq p_o(\mathbf{X}) < q_1\,]} - 1_{[\,0 \leq p_o(\mathbf{X}) < q_1'\,]})\,]\,|.$$

Then letting $[\, q_1, q_1' \,]$ refer to the real interval between the two values q_1 and q_1'.

$$\begin{aligned} |\, M(d_1; q_1) \;-\; M(d_1; q_1') \,| &\leq |\, (1 - 2d_1) \,| \int\limits_{[q_1, q_1']} f_p(r)\, dr \\ &\leq |\, (1 - 2d_1) \,| \times \sup_{r \in [q_1, q_1']} \{ f_p(r) \} \times |\, q_1 - q_1' \,|. \end{aligned}$$

So provided that the probability density function of the propensity score is bounded near q_{1o}, we can choose $k = |\, (1 - 2d_1) \,| \times \sup_{r \in [q_{1o} - \delta, q_{1o} + \delta]} \{ f_p(r) \}$. Then condition (A.5) is satisfied. Therefore, all the conditions of Theorem A.1 are satisfied and $\hat{d}_1$ is consistent.

Exactly the same argument can be applied to each estimated probability of being treated and in each stratum, showing that $\mathbf{d}_o$ is globally identifiable, and that when the probability density function of the propensity score is bounded near the population strata boundaries, $\hat{\mathbf{d}}$ is a consistent estimator of $\mathbf{d}_o$. The former result will be useful in demonstrating the consistency of the stratified treatment effect estimator.

A.2.3 Consistency of the stratified treatment effect estimator

We now show that when the strata boundaries and the probabilities of being treated and in each stratum are consistently estimated, the stratified treatment effect estimator, $\hat{\beta}^s$, is consistent. This estimator is obtained by solving the estimating equation $\sum_{i=1}^{n} \psi_1(Y_i, Z_i, \mathbf{X}_i; \beta^s, \hat{\mathbf{d}}, \hat{\mathbf{q}}) = 0$ where

$$\psi_1^{1\times 1}\, (Y_i, Z_i, \mathbf{X}_i\, ; \beta^s \hat{\mathbf{d}}, \hat{\mathbf{q}}) = \sum_{s=1}^{K} r_s \left\{ \frac{Y_i\, Z_i\, \hat{S}_s}{\hat{d}_s} - \frac{Y_i\, (1 - Z_i)\, \hat{S}_s}{r_s - \hat{d}_s} \right\} - \beta^s,$$

In order to apply Theorem A.1 to this estimator, we need to find a criteria function — a function that is maximised by the estimator $\hat{\beta}^s$. We can obtain this estimator by finding the value of β^s that maximises the criteria function

$$M_n(\beta^s; \hat{\mathbf{d}}, \hat{\mathbf{q}}) = -\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{s=1}^{K} r_s \left\{ \frac{Y_i\, Z_i\, \hat{S}_s}{\hat{d}_s} - \frac{Y_i\, (1 - Z_i)\, \hat{S}_s}{r_s - \hat{d}_s} \right\} - \beta^s \right)^2$$

We now merely need to verify the conditions of Theorem A.1. The criteria function $M_n(\beta^s; \mathbf{d}, \mathbf{q})$ is concave in β^s and so condition *(i)* of the theorem is satisfied. We have seen that both $\mathbf{q}_o$ and $\mathbf{d}_o$ are globally identifiable. We now need to show that β_o^s is

globally identifiable. By the law of large numbers, for any $\mathbf{d}$ and $\mathbf{q}$, as $n \to \infty$.

$$M_n(\beta^s; \mathbf{d}, \mathbf{q}) \xrightarrow{p} M(\beta^s; \mathbf{q}, \mathbf{d})$$

where

$$M(\beta^s; \mathbf{d}, \mathbf{q}) = -\mathbb{E}\left[\left(\sum_{s=1}^{K} r_s \left\{\frac{Y\,Z\,S_s}{d_s} - \frac{Y\,(1-Z)\,S_s}{r_s - d_s}\right\} - \beta^s\right)^2\right] \tag{A.6}$$

By expanding the expectation on the right-hand side, we see that the criteria function $M(\beta^s; \mathbf{d}_o, \mathbf{q}_o)$ is maximised at the point

$$\beta_o^s = \sum_{s=1}^{K} r_s \left\{\frac{\mathbb{E}[\,Y\,Z\,S_{so}\,]}{d_{so}} - \frac{\mathbb{E}[\,Y\,(1-Z)\,S_{so}\,]}{r_s - d_{so}}\right\}.$$

Remembering that $d_{so} = \mathbb{E}[\,ZS_{so}\,]$, this can be simplified to give

$$\beta_o^s = \sum_{s=1}^{K} r_s \left\{\mathbb{E}[\,Y \mid Z=1,\, S_{so}=1\,] - \mathbb{E}[\,Y \mid Z=0, S_{so}=1\,]\right\}.$$

This estimand β_o^s is unique provided that the population strata boundaries are globally identifiable. Then condition *(ii)* of Theorem A.1 is satisfied.

It only remains to verify condition *(iii*)* of Theorem A.1. In order to do this, we need to show that the Lipschitz condition holds for each nuisance parameter with respect to the criteria function $M(\beta^s; \mathbf{d}, \mathbf{q})$ in a neighbourhood of the population parameters. This is demonstrated for the first strata boundary, q_1, and the probability of being treated and in the first stratum, d_1.

Expanding the expectation on the right-hand side of (A.6), since $S_i\,S_j = 0$ for $i \neq j$, we see that

$$\begin{aligned} M(\beta^s; \mathbf{d}, \mathbf{q}) = &-\sum_{s=1}^{K} r_s^2\,\mathbb{E}\left[\frac{Y^2\,Z\,S_s}{d_s^2} + \frac{Y^2\,(1-Z)\,S_s}{(r_s - d_s)^2}\right] - (\beta^s)^2 \\ &+ 2\,\beta^s \sum_{s=1}^{K} r_s\,\mathbb{E}\left[\frac{Y\,Z\,S_s}{d_s} - \frac{Y\,(1-Z)\,S_s}{r_s - d_s}\right]. \end{aligned}$$

For the moment, we ignore all but the first term of this, and let

$$M^*(\beta^s; \mathbf{d}, \mathbf{q}) = -\sum_{s=1}^{K} r_s^2 \, \mathbb{E}\left[\frac{Y^2 \, Z \, S_s}{d_s^2}\right],$$

and show that the Lipschitz criteria holds for q_1 and d_1 with respect to this sub-function $M^*(\beta^s; \mathbf{d}, \mathbf{q})$. The whole Lipschitz condition can be demonstrated in the same way. Given $\mathbf{d}$, $\mathbf{q} = (q_1, q_2, ..., q_{(K-1)})^T$ and $\mathbf{q}' = (q_1', q_2, ... q_{(K-1)})^T$, where $q_1, q_1' \in [\, q_{1o} - \delta, q_{1o} + \delta \,]$,

$$\begin{aligned} |\, M^*(\beta^s; \mathbf{d}, \mathbf{q}) - M^*(\beta^s; \mathbf{d}, \mathbf{q}') \,| &= \left| \sum_{s=1}^{K} \frac{r_s^2}{d_s^2} \, \mathbb{E}\left[\, Y^2 \, Z \left\{ 1_{[\, q_{(s-1)} \le p_o(\mathbf{X}) < q_s \,]} - 1_{[\, q'_{(s-1)} \le p_o(\mathbf{X}) < q'_s \,]} \right\} \right] \right| \\ &= \frac{r_1^2}{d_1^2} \, |\, \mathbb{E}\left[\, Z \, Y^2 \left\{ 1_{[\, q_0 \le p_o(\mathbf{X}) < q_1 \,]} - 1_{[\, q_0 \le p_o(\mathbf{X}) < q_1' \,]} \right\} \right] |. \\ &\le \frac{r_1^2}{d_1^2} \int_{[q_1, q_1']} \mathbb{E}\left[\, Y^2 \,|\, Z = 1, \, p_o(\mathbf{X}) = r \,\right] f_p(r \,|\, Z = 1) \, dr, \end{aligned}$$

where, as before, $[\, q_1, q_1' \,]$ denotes the interval between the values q_1 and q_1'. Then, provided that $\mathbb{E}\left[\, Y^2 \,|\, Z = 1, \, p_o(\mathbf{X}) = r \,\right]$ and $f_p(r \,|\, Z = 1)$ are bounded near $r = q_{1o}$, we can pick

$$k = \frac{r_1^2}{d_1^2} \times \sup_{r \in [\, q_{1o} - \delta, q_{1o} + \delta \,]} \left\{ \mathbb{E}\left[\, Y^2 \,|\, Z = 1, \, p_o(\mathbf{X}) = r \,\right] \right\} \times \sup_{r \in [\, q_{1o} - \delta, q_{1o} + \delta \,]} \left\{ f_p(r \,|\, Z = 1) \right\},$$

in which case we have, as required,

$$|\, M^*(\beta^s; \mathbf{d}, \mathbf{q}) - M^*(\beta^s; \mathbf{d}, \mathbf{q}') \,| \le k \,|\, q_1 - q_1' \,|.$$

Note that since $\mathbb{P}(Z = 1)$ is non-zero, the condition that $f_p(r \,|\, Z = 1)$ is unbounded is equivalent to $f_p(r)$ being unbounded. We now show that the Lipschitz condition holds for d_1 with respect to the function $M^*(\beta^s; \mathbf{d}, \mathbf{q})$. Given $\mathbf{q}$, $\mathbf{d} = (d_1, d_2, ..., d_K)^T$ and $\mathbf{d}' = (d_1', d_2, ..., d_K)^T$, where $d_1, d_1' \in [\, d_{1o} - \delta, d_{1o} + \delta \,]$, We have

$$\begin{aligned} |\, M^*(\beta^s; \mathbf{d}, \mathbf{q}) - M^*(\beta^s; \mathbf{d}', \mathbf{q}) \,| &= \left| \sum_{s=1}^{K} r_s^2 \, \mathbb{E}\left[\, Y^2 \, Z \, S_s \,\right] \left(\frac{1}{d_s^2} - \frac{1}{(d_s')^2} \right) \right| \\ &= r_1^2 \left| \mathbb{E}\left[\, Y^2 \, Z \, S_1 \,\right] \left(\frac{1}{d_1^2} - \frac{1}{(d_1')^2} \right) \right|. \end{aligned}$$

By writing this as an integral, we find that the Lipschitz condition is satisfied provided that $\mathbb{E}\left[\, Y^2 \,|\, Z = 1, \, p_o(\mathbf{X}) = r \,\right]$ and $f_p(r)$ are bounded for all $r \in S_1$, and that d_{1o} is not equal to zero.

In this way, we can show that all the required Lipschitz conditions are satisfied provided that the functions $\mathbb{E}[Y^2 \mid Z = t, p_o(\mathbf{X}) = r]$, $\mathbb{E}[Y \mid Z = t, p_o(\mathbf{X}) = r]$ and $f_p(r)$ are bounded everywhere, for $t = 0, 1$, and d_s is not equal to zero or r_s for $s = 1, ..., K$. The same argument can be applied to each nuisance parameter in order to completely verify condition *(iii*)* of Theorem A.1. Then all the conditions have been satisfied and so the stratified treatment effect estimator is a consistent estimator of β_o^s. The implications of the conditions attached to this consistency are discussed in the main text (Section 3.2).

A.3 Asymptotic normality

We have established conditions under which the estimator $\hat{\boldsymbol{\theta}}$ is consistent, where $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T)$. We now consider the asymptotic distribution of $\hat{\boldsymbol{\theta}}$. We refer to Theorem 5.21 of van der Vaart [108, p.52]. This theorem concerns the asymptotic distribution of an estimator, $\hat{\boldsymbol{\theta}}$, obtained by solving an estimating equation $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{W}_i; \boldsymbol{\theta}) = 0$, where $\mathbf{W}$ represents a sample of data. The theorem gives conditions under which $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed.

Theorem A.2 *For each $\boldsymbol{\theta}$ in an open subset of Euclidean space, let $\mathbf{W} \rightarrow \boldsymbol{\psi}(\mathbf{W}; \boldsymbol{\theta})$ be a measurable vector-valued function such that, for every $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in a neighbourhood of $\boldsymbol{\theta}_o$ and some measurable function $\boldsymbol{\psi}'$ with $\mathbb{E}[\boldsymbol{\psi}'] < \infty$,*

$$|\boldsymbol{\psi}(\mathbf{W}; \boldsymbol{\theta}_1) - \boldsymbol{\psi}(\mathbf{W}; \boldsymbol{\theta}_2)| \leq \boldsymbol{\psi}' |\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|$$

Assume that $\mathbb{E}[|\boldsymbol{\psi}(\mathbf{W}; \boldsymbol{\theta}_o)|] < \infty$ and that the map $\boldsymbol{\theta} \rightarrow \mathbb{E}[\boldsymbol{\psi}(\mathbf{X}; \boldsymbol{\theta})]$ is differentiable at a zero $\boldsymbol{\theta}_o$, with a non-singular derivative matrix. If $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{W}; \hat{\boldsymbol{\theta}}) = o_p(n^{-1/2})$ and $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_o$, then

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = -\left(\frac{\partial}{\partial \boldsymbol{\theta}^T} \{\mathbb{E}[\boldsymbol{\psi}(\mathbf{W}; \boldsymbol{\theta})]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{W}_i; \boldsymbol{\theta}_o) + o_p(1).$$

In particular, the sequence $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$ is asymptotically normal with mean zero and covariance matrix

$$\left(\frac{\partial}{\partial \boldsymbol{\theta}^T} \{\mathbb{E}[\boldsymbol{\psi}(\mathbf{W}; \boldsymbol{\theta})]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \right)^{-1} \mathbb{E}[\boldsymbol{\psi}(\mathbf{W}; \boldsymbol{\theta}_o)\boldsymbol{\psi}^T(\mathbf{W}; \boldsymbol{\theta}_o)] \left(\frac{\partial}{\partial \boldsymbol{\theta}^T} \{\mathbb{E}[\boldsymbol{\psi}(\mathbf{W}; \boldsymbol{\theta})]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \right)^{-T}$$

◇

The Lipschitz condition in Theorem A.2 is in fact too strong a condition (see the discussion on p.53 and Lemma 19.24, of van der Vaart [108]). The Lipschitz condition can be replaced by the following conditions:

$$\begin{aligned} &(a) \text{ the functions } \mathbf{W} \rightarrow \boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta}) \text{ are a 'Donsker class'.} \\ &(b) \text{ the map } \boldsymbol{\theta} \rightarrow \boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta}) \text{ is continuous in probability.} \end{aligned} \tag{A.7}$$

We have already shown that the estimator $\hat{\boldsymbol{\theta}}$ is consistent. We have to verify the two conditions above and then show that the derivative matrix $\frac{\partial}{\partial \boldsymbol{\theta}^T} \{\mathbb{E}[\boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta})]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$ exists and is non-singular. Then we will have shown that $\hat{\boldsymbol{\theta}}$ is asymptotically normal.

We begin by verifying condition *(a)* of (A.7) above. Our estimating equation is defined by $\boldsymbol{\psi}^T = (\psi_1, \boldsymbol{\psi}_2^T, \boldsymbol{\psi}_3^T)$, where ψ_1, $\boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_3$ are defined in Section A.1, and the sample data are $\mathbf{W} = (Y, Z, \mathbf{X})$. We need to show that the functions $\mathbf{W} \rightarrow \boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta})$ are Donsker. This will be true if each component, $\mathbf{W} \rightarrow \boldsymbol{\psi}_j(\mathbf{W};\boldsymbol{\theta})$, is Donsker. We note that the map $\mathbf{W} \rightarrow \boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta})$ is Donsker, for $j = 1, ..., K-1$, by Example 19.6 of van der Vaart [108, p.271], which states that the class of all functions of the form $x \rightarrow 1_{(-\infty,t]}$ is Donsker. Example 19.20 of van der Vaart [108, p.277] states that if functions $f(x)$ and $g(x)$ are both Donsker then the function $f(x) + g(x)$ is Donsker and so is the function $f(x).g(x)$. Therefore, each map $\mathbf{W} \rightarrow \boldsymbol{\psi}_{2j}(\mathbf{W};\boldsymbol{\theta})$ is Donsker, for $j = 1, ..., K$. The same example also guarantees that the map $\mathbf{W} \rightarrow \psi_1(\mathbf{W};\boldsymbol{\theta})$ is Donsker. Therefore, the map $\mathbf{W} \rightarrow \boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta})$ is a Donsker class.

Condition *(b)* of (A.7) demands that the map $\boldsymbol{\theta} \rightarrow \boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta})$ should be continuous in probability. This is equivalent to each of the three sets of maps $\boldsymbol{\theta} \rightarrow \boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta})$, $\boldsymbol{\theta} \rightarrow \boldsymbol{\psi}_{2k}(\mathbf{W};\boldsymbol{\theta})$ and $\boldsymbol{\theta} \rightarrow \psi_1(\mathbf{W};\boldsymbol{\theta})$ being continuous in probability, for $j = 1, ..., K-1$, and $k = 1, ..., K$.

To show that the map $\boldsymbol{\theta} \rightarrow \boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta})$ is continuous in probability we need to show that for $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ defined by $\boldsymbol{\theta}^T = (\beta^s, \mathbf{q}^T, \mathbf{d}^T)$ and $\boldsymbol{\theta}'^T = ((\beta^s)', \mathbf{q}'^T, \mathbf{d}'^T)$,

$$\mathbb{P}(|\,\boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta}) - \boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta}')\,| \geq \epsilon) \rightarrow 0, \quad \text{as} \quad \boldsymbol{\theta} \rightarrow \boldsymbol{\theta}',$$

or, equivalently, that

$$\mathbb{E}[\,|\,\boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta}) - \boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta}')\,|\,] \rightarrow 0, \quad \text{as} \quad \boldsymbol{\theta} \rightarrow \boldsymbol{\theta}'.$$

Using the definition of ψ_3 given in Section A.1,

$$\mathbb{E}\left[\,|\,\psi_{3j}(\mathbf{W};\boldsymbol{\theta}) - \psi_{3j}(\mathbf{W};\boldsymbol{\theta}')\,|\,\right] \;=\; \mathbb{E}\left[\,|1_{[q_{(j-1)}\le p_o(\mathbf{X})<q_j]} - 1_{[q'_{(j-1)}\le p_o(\mathbf{X})<q'_j]}|\,\right].$$

Using the triangle inequality,

$$\begin{aligned}\mathbb{E}\left[\,|\,\psi_{3j}(\mathbf{W};\boldsymbol{\theta}) - \psi_{3j}(\mathbf{W};\boldsymbol{\theta}')\,|\,\right] \le\;& \mathbb{E}\left[\,|1_{[q_{(j-1)}\le p_o(\mathbf{X})<q_j]} - 1_{[q_{(j-1)}\le p_o(\mathbf{X})<q'_j]}|\,\right] \\ &+ \mathbb{E}\left[\,|1_{[q_{(j-1)}\le p_o(\mathbf{X})<q'_j]} - 1_{[q'_{(j-1)}\le p_o(\mathbf{X})<q'_j]}|\,\right]\end{aligned}$$

We can write this in integral form as

$$\mathbb{E}\left[\,|\,\psi_{3j}(\mathbf{W};\boldsymbol{\theta}) - \psi_{3j}(\mathbf{W};\boldsymbol{\theta}')\,|\,\right] \le \int_{[q_j,q_{j'}]} f_p(p)\;dp \;+\; \int_{[q_{(j-1)},q'_{(j-1)}]} f_p(p)\;dp.$$

Then,

$$\begin{aligned}&\mathbb{E}\left[\,|\,\psi_{3j}(\mathbf{W};\boldsymbol{\theta}) - \psi_{3j}(\mathbf{W};\boldsymbol{\theta}')\,|\,\right] \\ &\le |\,q_j - q'_j\,| \times \sup_{r\in[q_j,q'_j]}\{f_p(p)\} + |\,q_{(j-1)} - q'_{(j-1)}\,| \times \sup_{r\in[q_{(j-1)},q'_{(j-1)}]}\{f_p(p)\} \;\to\; 0\end{aligned}$$

as $q_{(j-1)} \to q_{(j-1)'}$ and $q_j \to q_{j'}$, provided that probability density function of the propensity score is bounded. Then the map $\boldsymbol{\theta} \to \psi_{3j}(\mathbf{W};\boldsymbol{\theta})$ is continuous in probability.

We now turn to the map $\boldsymbol{\theta} \to \psi_{2k}(\mathbf{W};\boldsymbol{\theta})$. Using the definition of ψ_2 given in Section A.1,

$$\begin{aligned}&\mathbb{E}\left[\,|\,\psi_{2k}(\mathbf{W};\boldsymbol{\theta}) - \psi_{2k}(\mathbf{W};\boldsymbol{\theta}')\,|\,\right] \\ =\;& \mathbb{E}\left[\,|\,\{Z\,1_{[q_{(k-1)}\le p_o(\mathbf{X})<q_k]} - d_k\} - \{Z\,_{[q'_{(k-1)}\le p_o(\mathbf{X})<q'_k]} - d'_k\}\,|\,\right].\end{aligned}$$

And applying the triangle equality gives,

$$\begin{aligned}&\mathbb{E}\left[\,|\,\psi_{2k}(\mathbf{W};\boldsymbol{\theta}) - \psi_{2k}(\mathbf{W};\boldsymbol{\theta}')\,|\,\right] \\ \le\;& \mathbb{E}\left[\,|1_{[q_{(k-1)}\le p_o(\mathbf{X})<q_k]} - 1_{[q'_{(k-1)}\le p_o(\mathbf{X})<q'_k]}|\,\right] + |\,d_k - d'_k\,| \to 0,\end{aligned}$$

as $q_{(k-1)} \to q_{(k-1)'}$, $q_k \to q_{k'}$, and $d_k \to d'_k$, provided again that the probability distribution function of the propensity score is bounded. Then the map $\boldsymbol{\theta} \to \psi_{2k}(\mathbf{W};\boldsymbol{\theta})$

is continuous in probability.

We finally need to show that the map $\boldsymbol{\theta} \rightarrow \psi_1(\mathbf{W}; \boldsymbol{\theta})$ is continuous in probability. This is the map

$$\boldsymbol{\theta} \rightarrow \sum_{s=1}^{K} r_s \left\{ \frac{Y\,Z\,S_s}{d_s} - \frac{Y\,(1-Z)\,S_s}{r_s - d_s} \right\} - \beta^s.$$

Using the triangle equality, we see that this map will be continuous in probability provided that the expected outcome $\mathbb{E}\,[\,Y \mid Z = t,\, S_s = s\,]$ is bounded and that for each $s = 1, ..., K$,

$$\begin{aligned} \mathbb{E}\left[\,\left|\frac{S_s}{d_s} - \frac{S'_s}{d'_s}\right|\,\right] &\rightarrow 0 \\ \mathbb{E}\left[\,\left|\frac{S_s}{r_s - d_s} - \frac{S'_s}{r_s - d'_s}\right|\,\right] &\rightarrow 0, \end{aligned} \tag{A.8}$$

as $\mathbf{q} \rightarrow \mathbf{q}'$ and $\mathbf{d} \rightarrow \mathbf{d}'$. Now

$$\mathbb{E}\left[\,\left|\frac{S_s}{d_s} - \frac{S'_s}{d'_s}\right|\,\right] = \mathbb{E}\left[\,\left|\frac{1_{[q_{(s-1)} \leq p_o(\mathbf{X}) < q_s]}}{d_s} - \frac{1_{[q_{(s-1)'} \leq p_o(\mathbf{X}) < q'_s]}}{d'_s}\right|\,\right] \leq e_1 + e_2 + e_3,$$

where

$$\begin{aligned} e_1 &= \frac{1}{d_s}\,\mathbb{E}\,[\,|1_{[q_{(s-1)} \leq p_o(\mathbf{X}) < q_s]} - 1_{[q_{(s-1)'} \leq p_o(\mathbf{X}) < q_s]}|\,] \\ e_2 &= \frac{1}{d_s}\,\mathbb{E}\,[\,|1_{[q_{(s-1)'} \leq p_o(\mathbf{X}) < q_s]} - 1_{[q_{(s-1)'} \leq p_o(\mathbf{X}) < q'_s]}|\,] \\ e_3 &= \mathbb{E}\left[\,\left|\frac{1_{[q_{(s-1)'} \leq p_o(\mathbf{X}) < q'_s]}}{d_s} - \frac{1_{[q_{(s-1)'} \leq p_o(\mathbf{X}) < q'_s]}}{d'_s}\right|\,\right]. \end{aligned}$$

We have already seen that when the probability density function of the propensity score is bounded, $e_1 \rightarrow 0$ and $e_2 \rightarrow 0$ as $\mathbf{q} \rightarrow \mathbf{q}'$. Also, $e_3 \rightarrow 0$ as $\mathbf{d} \rightarrow \mathbf{d}'$ provided that d_s and d'_s are non-zero. If also, d_s and d'_s are not equal to r_s, then the map $\boldsymbol{\theta} \rightarrow \psi_1(\mathbf{W}; \boldsymbol{\theta})$ is continuous in probability.

In this way, we see that if the probability density function of the propensity score is continuous and the population probabilities d_{so} are not equal to zero or r_s for $s = 1, ..., K$, then the map $\boldsymbol{\theta} \rightarrow \psi(\mathbf{W}; \boldsymbol{\theta})$ is continuous in probability.

We have now verified each condition of Theorem A.2 other than the requirement that the derivative matrix $\frac{\partial}{\partial \boldsymbol{\theta}^T}\{\mathbb{E}[\boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta})]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$ should exist and be non-singular. This matrix is calculated in Section A.4, and conditions under which it exists and is invertible are stated. When the inverse of this derivative matrix exists, the probability density function of the propensity score is continuous, and the population probabilities d_{so} are not equal to zero or r_s for $s = 1, ..., K$, then all the conditions of Theorem A.2 are satisfied and the estimator $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed.

A.4 Asymptotic variance

We have now established conditions under which the estimator $\hat{\boldsymbol{\theta}}$, defined by $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T)$, is consistent and asymptotically normally distributed. We now use its asymptotic distribution to calculate the asymptotic variance of the stratified treatment effect estimator, $\hat{\beta}^s$. We denote this by $\mathbb{V}_k[\hat{\beta}^s]$ where the 'k' subscript refers to the propensity score being known.

A.4.1 M-estimation theory

The general M-estimation theory outlined in the main text (Section 3.1.2) shows that if we let

$$\begin{aligned} A &= \mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}^T}\{\boldsymbol{\psi}(Y, Z, \mathbf{X};\, \boldsymbol{\theta})\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\right] \\ B &= \mathbb{E}[\,\boldsymbol{\psi}(Y, Z, \mathbf{X};\, \boldsymbol{\theta}_o)\;\boldsymbol{\psi}^T(Y, Z, \mathbf{X};\, \boldsymbol{\theta}_o)\,], \end{aligned} \tag{A.9}$$

then the large-sample covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\,\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-\mathbf{T}}. \tag{A.10}$$

We furthermore stated that when the function $\boldsymbol{\psi}(Y, Z, \mathbf{X};\, \boldsymbol{\theta})$ is not differentiable with respect to $\boldsymbol{\theta}$, as in our example, the order of differentiation and expectation can be exchanged and so

$$A = -\frac{\partial}{\partial \boldsymbol{\theta}^T}\{\mathbb{E}[\,\boldsymbol{\psi}(Y, Z, \mathbf{X};\, \boldsymbol{\theta})\,]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}. \tag{A.11}$$

Although we do not list the regularity conditions under which this exchange is valid, the application of Theorem A.2 shows that it is valid in our problem. Therefore, when a component of $\boldsymbol{\psi}(Y, Z, \mathbf{X};\, \boldsymbol{\theta})$ is not differentiable with respect to $\boldsymbol{\theta}$ we use (A.11).

For components of $\psi(Y, Z, \mathbf{X}; \boldsymbol{\theta})$ that are differentiable with respect to $\boldsymbol{\theta}$, however, since the two versions of A are then identical, we usually use (A.9), since this is often more convenient.

Allowing for definition (A.11) to be used if necessary, (A.10) is then equal to the variance given at the end of Theorem A.2, and the large-sample variance of the stratified treatment effect estimator, $\hat{\beta}^s$ is

$$\mathbb{V}_k\left[\,\hat{\beta}^s\,\right] = \mathbb{V}\left[\,(\hat{\boldsymbol{\theta}})_1\,\right] = \frac{1}{n}\left(\mathbf{A^{-1}B A^{-T}}\right)_{11}.$$

We have partitioned the unknown parameter, $\boldsymbol{\theta}$, into three components as follows, $\boldsymbol{\theta}^T = (\beta^s, \mathbf{d}^T, \mathbf{q}^T)$. The matrices $A^{2K\times 2K}$ and $B^{2K\times 2K}$ can be partitioned in the same way. Then,

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

where for $j = 1, 2, 3$,

$$a_{j1} = -\frac{\partial}{\partial \beta^s}\left\{\mathbb{E}\left[\,\psi_j\,\right]\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \qquad a_{j2} = -\frac{\partial}{\partial \mathbf{d}^T}\left\{\mathbb{E}\left[\,\psi_j\,\right]\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$a_{j3} = -\frac{\partial}{\partial \mathbf{q}^T}\left\{\mathbb{E}\left[\,\psi_j\,\right]\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

These sub-matrices of A have the following dimensions:

$$Dim = \begin{pmatrix} 1\times 1 & 1\times K & 1\times(K-1) \\ K\times 1 & K\times K & K\times(K-1) \\ (K-1)\times 1 & (K-1)\times K & (K-1)\times(K-1) \end{pmatrix} \tag{A.12}$$

We can simplify the matrix A immediately. Since the functions ψ_2 and ψ_3 do not contain β^s, differentiating with respect to β^s results in zero and so the sub-matrices a_{21} and a_{31} are zero matrices. Similarly, the function ψ_3 does not contain $\mathbf{d}$, so a_{32} is a zero matrix. Differentiating ψ_1 with respect to β^s gives -1 and similarly, differentiating ψ_2 with respect to $\mathbf{d}$ gives $-I$, where I is the identity matrix. Then

the matrix A is

$$A = \begin{pmatrix} 1 & a_{12} & a_{13} \\ 0 & I & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix},$$

which has inverse

$$A^{-1} = \begin{pmatrix} 1 & -a_{12} & (a_{12}a_{23} - a_{13})a_{33}^{-1} \\ 0 & I & -a_{23}a_{33}^{-1} \\ 0 & 0 & a_{33}^{-1} \end{pmatrix}.$$

The conditions under which the sub-matrices a_{12}, a_{13}, a_{23}, and a_{33}^{-1} exist are investigated later in this section. The existence of these sub-matrices and hence the existence of the inverse A^{-1} is one of the conditions for asymptotic normality of $\hat{\boldsymbol{\theta}}$.

Partitioning the matrix B in a similar fashion gives

$$B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix},$$

where $b_{jk} = \mathbb{E}[\,\boldsymbol{\psi}_j(\boldsymbol{\theta}_o)\;\boldsymbol{\psi}_k^T(\boldsymbol{\theta}_o)\,]$, for $j,k = 1,2,3$. The dimensions of these sub-matrices are also given by (A.12).

We have seen that $n\,\mathbb{V}_k[\,\hat{\beta}^s\,] = (A^{-1}BA^{-T})_{11}$. Multiplying out $A^{-1}BA^{-T}$ and taking the $(1,1)^{th}$ component gives

$$\begin{aligned} n\,\mathbb{V}_k[\,\hat{\beta}^s\,] = {} & b_{11} - b_{12}a_{12}^T + b_{13}a_{33}^{-T}(a_{12}a_{23} - a_{13})^T \\ & - a_{12}(b_{21} - b_{22}a_{12}^T + b_{23}a_{33}^{-T}(a_{12}a_{23} - a_{13})^T \\ & + (a_{12}a_{23} - a_{13})a_{33}^{-1}(b_{31} - b_{32}a_{12}^T + b_{33}a_{33}^{-T}(a_{12}a_{23} - a_{13})^T). \end{aligned}$$

Remembering that the matrix B is symmetric, so for example $b_{13} = b_{31}^T$, we can write this as,

$$\begin{aligned} n\,\mathbb{V}_k[\,\hat{\beta}^s\,] = {} & b_{11} - 2b_{12}a_{12}^T + a_{12}b_{22}a_{12}^T \\ & + (a_{12}a_{23} - a_{13})a_{33}^{-1}\{2b_{31} - 2b_{32}a_{12}^T + b_{33}a_{33}^{-T}(a_{12}a_{23} - a_{13})^T\}. \end{aligned} \qquad \text{(A.13)}$$

The sub-matrices contained in this formula are calculated in Sections A.4.2 and A.4.3, and are substituted into the variance equation (A.13). The variance $n\,\mathbb{V}[\,\hat{\beta}^s\,]$ is then calculated by direct matrix multiplication.

A.4.2 The matrix A

We now calculate the sub-matrices of A of interest, that is a_{12}, a_{13}, a_{23} and a_{33}^{-1}. Since subjects are sampled independently from the population we drop the subject subscripts in the following calculations in order to simplify the presentation.

The sub-matrix $\mathbf{a_{12}}$ The function ψ_1 is differentiable with respect to $\mathbf{d}$ so we define, for $j = 1, .., K$,

$$\begin{aligned}(a_{12})_j &= -\mathbb{E}\left[\frac{\partial}{\partial d_j}\ \{\,\psi_1\,\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\right] \\ &= -\mathbb{E}\left[\frac{\partial}{\partial d_j}\left\{\sum_{s=1}^{K} r_s\left(\frac{Y\,Z\,S_s}{d_s} - \frac{Y\,(1-Z)\,S_s}{r_s - d_s}\right) - \beta^s\right\}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\right].\end{aligned}$$

Differentiating this with respect to d_j and setting all parameters to their true values gives

$$(a_{12})_j = -\mathbb{E}\left[-\frac{r_j\,Y\,Z\,S_{jo}}{(d_{jo})^2} - \frac{r_j\,Y\,(1-Z)\,S_{jo}}{(r_j - d_{jo})^2}\right]$$

We now use the equality $\mathbb{E}[\,Y\,Z\,S_s\,] = \mathbb{E}[\,Y \mid Z = 1,\ S_s = 1\,]\,\mathbb{E}[\,ZS_s\,]$. Then,

$$\begin{aligned}(a_{12})_j = &\frac{r_j}{(d_{jo})^2}\,\mathbb{E}[\,Y \mid Z = 1,\ S_{jo} = 1\,]\,\mathbb{E}[\,Z\,S_{jo}\,] \\ &+ \frac{r_j}{(r_j - d_{jo})^2}\,\mathbb{E}\,[\,Y \mid Z = 0,\ S_{jo} = 1\,]\,\mathbb{E}[\,(1-Z)S_{jo}\,].\end{aligned}$$

This can be simplified using the fact that $\mathbb{E}[\,\boldsymbol{\psi}(\boldsymbol{\theta}_o)\,] = 0$, and so $\mathbb{E}[\,ZS_{jo}\,] = d_{jo}$. Then, for $j = 1, 2, ..., K$, we define a_{12} as

$$(a_{12})_j = \frac{r_j}{d_{jo}}\,\mathbb{E}[\,Y \mid Z = 1,\ S_{jo} = 1\,] + \frac{r_j}{r_j - d_{jo}}\,\mathbb{E}[\,Y \mid Z = 0,\ S_{jo} = 1\,].$$

The sub-matrix $\mathbf{a_{13}}$ Since the function ψ_1 is not differentiable with respect to $\mathbf{q}$ we

swap the order of differentiation and expectation and define, for $k = 1, .., K-1$.

$$(a_{13})_k = -\frac{\partial}{\partial q_k} \{\mathbb{E}[\,(\psi_1)\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$= -\frac{\partial}{\partial q_k} \left\{ \sum_{s=1}^{K} r_s \,\mathbb{E}\left[\frac{Y\,Z\,S_s}{d_s} - \frac{Y\,(1-Z)\,S_s}{r_s - d_s} - \beta^s \right] \right\} \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

This formula can be simplified further. We can write

$$-\frac{\partial}{\partial q_k}\,\mathbb{E}[\,Y\,Z\,S_s\,]|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -\frac{\partial}{\partial q_k}\,\{\mathbb{E}[\,Y \mid Z=1, S_s=1\,]\,\mathbb{E}[\,ZS_s\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

Differentiating this with respect to q_k using the product rule gives,

$$-\frac{\partial}{\partial q_k}\,\mathbb{E}[\,Y\,Z\,S_s\,]|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -\mathbb{E}[\,Y \mid Z=1, S_{so}=1\,]\,\frac{\partial}{\partial q_k}\,\{\mathbb{E}[\,ZS_s\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$-\,\mathbb{E}[\,ZS_{so}\,]\,\frac{\partial}{\partial q_k}\,\{\mathbb{E}[\,Y \mid Z=1, S_s=1\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

Since $\mathbb{E}[\,\psi(\boldsymbol{\theta}_o)\,] = 0$, we have $\mathbb{E}[\,Z\,S_{so}\,] = d_{so}$. Then,

$$-\frac{\partial}{\partial q_k}\,\mathbb{E}[\,Y\,Z\,S_s\,]|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -\mathbb{E}[\,Y \mid Z=1, S_{so}=1\,]\,\frac{\partial}{\partial q_k}\,\{\mathbb{E}[\,ZS_s\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$-\,d_{so}\,\frac{\partial}{\partial q_k}\,\{\mathbb{E}[\,Y \mid Z=1, S_s=1\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

Substituting this expression, and a similar one for $\mathbb{E}[\,Y\,(1-Z)\,S_s\,]$, we define, for $k = 1, 2, ..., K-1$,

$$(a_{13})_k = -\sum_{s=1}^{K} \frac{r_s}{d_{so}}\,\mathbb{E}[\,Y \mid Z=1, S_{so}=1\,]\,\frac{\partial}{\partial q_k}\,\{\mathbb{E}[\,Z\,S_s\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$-\sum_{s=1}^{K} r_s\,\frac{\partial}{\partial q_k}\,\{\mathbb{E}[\,Y \mid Z=1,\,S_s=1\,] - \mathbb{E}[\,Y \mid Z=0,\,S_s=1\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$+\sum_{s=1}^{K} \frac{r_s}{r_s - d_{so}}\,\mathbb{E}[\,Y \mid Z=0,\,S_{so}=1\,]\,\frac{\partial}{\partial q_k}\,\{\mathbb{E}[\,(1-Z)\,S_s\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

The sub-matrix $\mathbf{a_{23}}$ Again, we switch the order of expectation and differentiation and define, for $j = 1, ..., K$, and $k = 1, .., K-1$,

$$(a_{23})_{jk} = -\frac{\partial}{\partial q_k}\,\{\mathbb{E}[\,\psi_{2j}\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$= -\frac{\partial}{\partial q_k}\,\{\mathbb{E}[\,Z\,S_j\,]\}\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

The sub-matrix $\mathbf{a_{33}}$ For $j, k = 1, 2, ..., K-1$,

$$\begin{aligned}(a_{33})_{jk} &= -\frac{\partial}{\partial q_k}\left\{\mathbb{E}\left[\psi_{3j}\right]\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \\ &= -\frac{\partial}{\partial q_k}\left\{\mathbb{E}\left[S_j\right]\right\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.\end{aligned}$$

Remembering that $S_j = 1_{[q_{j-1} \le p_o(\mathbf{X}) < q_j]}$, and $f_p(\cdot)$ denotes the probability density function of the propensity score,

$$-\frac{\partial}{\partial q_k}\left\{\mathbb{E}\left[S_j\right]\right\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -\frac{\partial}{\partial q_k}\int_{q_{j-1}}^{q_j} f_p(p)\,dp\,\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \begin{cases} -f_p(q_{ko}) & \text{if } k = j \\ f_p(q_{ko}) & \text{if } k = j-1 \\ 0 & \text{otherwise.}\end{cases}$$

Therefore, the matrix a_{33} is

$$a_{33} = \begin{pmatrix} -f_p(q_{1o}) & 0 & 0 & \cdots & 0 \\ f_p(q_{1o}) & -f_p(q_{2o}) & 0 & \cdots & 0 \\ 0 & f_p(q_{2o}) & -f_p(q_{3o}) & 0 & 0 \\ \vdots & & & \ddots & 0 \\ 0 & \cdots & & f_p(q_{(K-2)o}) & -f_p(q_{(K-1)o}) \end{pmatrix},$$

which has inverse

$$(a_{33})^{-1} = -\begin{pmatrix} f_p(q_{1o})^{-1} & 0 & 0 & \cdots & 0 \\ f_p(q_{2o})^{-1} & f_p(q_{2o})^{-1} & 0 & & 0 \\ f_p(q_{3o})^{-1} & f_p(q_{3o})^{-1} & f_p(q_{3o})^{-1} & & 0 \\ \vdots & & & \ddots & \\ f_p(q_{(K-1)o})^{-1} & & & & f_p(q_{(K-1)o})^{-1} \end{pmatrix}.$$

Conditions under which A^{-1} exists

The inverse of the matrix A will exist if all its components exist. Inspection of the sub-matrices calculated in this section shows that this will happen provided that:

- the probabilities d_s are not equal to 0 or r_s for $s = 1, ..., K$;
- the probability density function of the propensity score is non-zero at each of the population strata boundaries;

- the following derivatives exist, for $t = 0, 1$, $j = 1, ..., K-1$ and $s = 1, ..., K$:

$$\frac{\partial}{\partial q_k} \{\mathbb{E}[Y \mid Z = t, S_s = 1]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}, \quad \frac{\partial}{\partial q_k} \{\mathbb{E}[Z S_s]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}, \quad \frac{\partial}{\partial q_k} \{\mathbb{E}[S_s]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

By writing these derivatives as integrals over the propensity score, and appealing to the fundamental theorem of calculus, we see that these derivatives exist provided that both the probability density function of the propensity score and $\mathbb{E}[Y \mid Z = t, p_o(\mathbf{X}) = p]$ are continuous near the population strata boundaries, for $t = 0, 1$.

A.4.3 The matrix B

We now calculate the sub-matrices of B of interest, that is $b_{11}, b_{12}, b_{13}, b_{22}, b_{23}$ and b_{33}, using the formula $b_{jk} = \mathbb{E}[\boldsymbol{\psi}_j(\boldsymbol{\theta}_o)\, \boldsymbol{\psi}_k^T(\boldsymbol{\theta}_o)]$, for $j, k = 1, 2, 3$, where $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_3$ are defined in Section A.1.

The sub-matrix $\mathbf{b_{11}}$ This is defined as

$$\begin{aligned} b_{11} &= \mathbb{E}[\boldsymbol{\psi}_1(\boldsymbol{\theta}_o)^2] \\ &= \mathbb{E}\left[\left(Y Z \sum_{s=1}^{K} \frac{r_s S_{so}}{d_{so}} - Y(1-Z) \sum_{s=1}^{K} \frac{r_s S_{so}}{r_s - d_{so}} - \beta_o^s\right)^2\right] \\ &= \sum_{s=1}^{K} r_s \left\{\mathbb{E}\left[\frac{Y^2 Z S_{so}}{d_{so}} + \frac{Y^2 (1-Z) S_{so}}{r_s - d_{so}}\right] - 2\beta_o^s \mathbb{E}\left[\frac{Y Z S_{so}}{d_{so}} - \frac{Y(1-Z) S_{so}}{r_s - d_{so}}\right]\right\} + (\beta_o^s)^2 \end{aligned}$$

Remembering that $\mathbb{E}[Z S_{so}] = d_{so}$ and $\mathbb{E}[(1-Z) S_{so}] = r_s - d_{so}$, this can be simplified to

$$\begin{aligned} b_{11} = \sum_{s=1}^{K} r_s^2 \left\{\frac{\mathbb{E}[Y^2 \mid Z = 1, S_{so} = 1]}{d_{so}} + \frac{\mathbb{E}[Y^2 \mid Z = 0, S_{so} = 1]}{r_s - d_{so}}\right\} \\ - 2\beta_o^s \sum_{s=1}^{K} r_s \{\mathbb{E}[Y \mid Z = 1, S_{so} = 1] - \mathbb{E}[Y \mid Z = 0, S_{so} = 1]\} + (\beta_o^s)^2. \end{aligned}$$

And since the population stratified treatment effect is defined as

$$\beta_o^s = \sum_{s=1}^{K} r_s \{\mathbb{E}[Y \mid Z = 1, S_{so} = 1] - \mathbb{E}[Y \mid Z = 0, S_{so} = 1]\},$$

we have

$$b_{11} = \sum_{s=1}^{K} r_s^2 \left\{ \frac{\mathbb{E}[Y^2 \mid Z=1,\, S_{so}=1]}{d_{so}} + \frac{\mathbb{E}[Y^2 \mid Z=0,\, S_{so}=1]}{r_s - d_{so}} \right\} - (\beta_o^s)^2.$$

The sub-matrix $\mathbf{b_{12}}$ For $j = 1, 2, ..., K$,

$$\begin{aligned}(b_{12})_j &= \mathbb{E}\,[\,\psi_1(\boldsymbol{\theta}_o)\;\boldsymbol{\psi}_{2j}^T(\boldsymbol{\theta}_o)\,] \\ &= \mathbb{E}\left[\left(Y\,Z\sum_{s=1}^{K}\frac{r_s S_{so}}{d_{so}} - Y\,(1-Z)\sum_{s=1}^{K}\frac{r_s S_{so}}{r_s - d_{so}} - \beta_o^s\right)(Z\,S_{jo} - d_{jo})\right].\end{aligned}$$

Now $\mathbb{E}\,[\,\psi_1(\boldsymbol{\theta}_o)\,d_{jo}\,] = d_{jo}\,\mathbb{E}\,[\,\psi_1(\boldsymbol{\theta}_o)\,] = 0$, since $\mathbb{E}\,[\boldsymbol{\psi}(\boldsymbol{\theta}_o)] = 0$. Therefore,

$$(b_{12})_j = \mathbb{E}\left[\left(Y\,Z\sum_{s=1}^{K}\frac{r_s S_{so}}{d_{so}} - Y\,(1-Z)\sum_{s=1}^{K}\frac{r_s S_{so}}{r_s - d_{so}} - \beta_o^s\right)\quad Z\,S_{jo}\right].$$

This is equal to

$$(b_{12})_j = \mathbb{E}\left[\frac{r_j\,Y\,Z\,S_{jo}}{d_{jo}} - \beta_o^s Z\,S_{jo}\right] = r_j\,\mathbb{E}\,[Y \mid Z=1,\, S_{jo}=1\,] - \beta_o^s\,d_{jo}.$$

The sub-matrix $\mathbf{b_{13}}$ For $j = 1, 2, ..., K-1$,

$$\begin{aligned}(b_{13})_j &= \mathbb{E}\,[\,\psi_1(\boldsymbol{\theta}_o)\;\boldsymbol{\psi}_{3j}^T(\boldsymbol{\theta}_o)\,] \\ &= \mathbb{E}\left[\left(Y\,Z\sum_{s=1}^{K}\frac{r_s S_{so}}{d_{so}} - Y\,(1-Z)\sum_{s=1}^{K}\frac{r_s S_{so}}{r_s - d_{so}} - \beta_o^s\right)(S_{jo} - r_j)\right] \\ &= r_j\,\{\mathbb{E}\,[Y \mid Z=1,\, S_{jo}=1\,] - \;\mathbb{E}\,[Y \mid Z=0,\, S_{jo}=1\,] - \beta_o^s\}.\end{aligned}$$

The sub-matrix $\mathbf{b_{22}}$ For $j, k = 1, ..., K$,

$$\begin{aligned}(b_{22})_{jk} &= \mathbb{E}\,[\,\boldsymbol{\psi}_{2j}(\boldsymbol{\theta}_o)\;\boldsymbol{\psi}_{2k}^T(\boldsymbol{\theta}_o)\,] = \mathbb{E}\,[\,(Z\;S_{jo} - d_{jo})\;(Z\;S_{ko} - d_{ko})\,] \\ &= \begin{cases} d_{jo}(1-d_{jo}) & \text{if } j = k, \\ -d_{jo}\;d_{ko} & \text{if } j \neq k. \end{cases}\end{aligned}$$

The sub-matrix $\mathbf{b_{33}}$ For $j, k = 1, 2, ..., K-1$,

$$(b_{33})_{jk} = \mathbb{E}\left[\,\psi_{3j}(\boldsymbol{\theta}_o)\;\psi_{3k}^T(\boldsymbol{\theta}_o)\,\right] = \mathbb{E}\left[\,(S_{jo} - r_j)\;(S_{ko} - r_k)\,\right]$$
$$= \begin{cases} r_j(1 - r_j) & \text{if } j = k, \\ -r_j\, r_k & \text{if } j \neq k. \end{cases}$$

The sub-matrix $\mathbf{b_{23}}$ For $j = 1, 2..., K$ and $k = 1, 2, ..., K-1$,

$$(b_{23})_{jk} = \mathbb{E}\left[\,\psi_{2j}(\boldsymbol{\theta}_o)\;\psi_{3k}^T(\boldsymbol{\theta}_o)\,\right] = \mathbb{E}\left[\,(ZS_{jo} - d_{jo})\;(S_{ko} - r_k)\,\right]$$
$$= \begin{cases} d_{jo}(1 - r_j) & \text{if } j = k, \\ -d_{jo}\, r_k & \text{if } j \neq k. \end{cases}$$

A.4.4 Variance of the stratified treatment effect estimator

We have already seen that

$$n\,\mathbb{V}_k[\,\hat{\beta}^s\,] = b_{11} - 2b_{12}a_{12}^T + a_{12}b_{22}a_{12}^T$$
$$+ (a_{12}a_{23} - a_{13})a_{33}^{-1}\{2b_{31} - 2b_{32}a_{12}^T + b_{33}a_{33}^{-T}(a_{12}a_{23} - a_{13})^T\}.$$

The sub-matrices involved in this formula have been calculated. Some rather lengthy direct matrix multiplication of the right-hand side of this equation shows that

$$\mathbb{V}_k[\,\hat{\beta}^s\,] = V_1 + V_2,$$

where, remembering that $S_{so} = 1_{[q_{(s-1)o} \leq p_o(\mathbf{X}) < q_s]}$ is an indicator for the s^{th} population stratum,

$$V_1 = \sum_{s=1}^{K} \frac{r_s^2}{n} \left\{ \frac{\mathbb{V}\,[\,Y \mid Z = 1,\, S_{so} = 1\,]}{d_{so}} + \frac{\mathbb{V}\,[\,Y \mid Z = 0,\, S_{so} = 1\,]}{r_s - d_{so}} \right\}$$

$$V_2 = \frac{1}{n} \left.\frac{\partial \beta^*}{\partial \mathbf{q}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} (n\ \mathrm{Cov}[\hat{\mathbf{q}}]) \left.\frac{\partial \beta^*}{\partial \mathbf{q}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o},$$

where $(n\,\mathrm{Cov}[\hat{\mathbf{q}}])$ is a $(K-1)\times(K-1)$ matrix representing the asymptotic covariance matrix of the estimated strata boundaries, defined for $j,k=1,....K-1,\ j\geq k$, as

$$(n\ \mathrm{Cov}[\,\hat{\mathbf{q}}\,])_{jk}=\frac{\mathbb{P}(p_o(\mathbf{X})>q_{jo})\ \mathbb{P}(p_o(\mathbf{X})<q_{ko})}{f_p(q_{jo})\ f_p(q_{ko})},$$

where $f_p(\,\cdot\,)$ is the probability density function of the propensity score and $\left.\frac{\partial\beta^*}{\partial\mathbf{q}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$ is a $1\times(K-1)$ vector with

$$\beta^*=\sum_{s=1}^{K}r_s\{\mathbb{E}\,[\,Y\mid Z=1,S_s=1\,]-\mathbb{E}\,[\,Y\mid Z=0,S_s=1\,]\},$$

which is equal to the 'true' value of $\hat{\beta}^s$, β_o^s, but seen as a function of the strata boundaries, $\mathbf{q}$, rather than evaluated at the population strata boundaries. If we were to replace S_s by S_{so}, then β^* would be equal to β_o^s, our parameter of interest.

We have therefore now derived conditions under which $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal, and we have calculated its asymptotic variance. The implications of this variance formula and the variance components are discussed in the main text, (Section 3.4).

Appendix B

Proof of Theorem 3.2

B.1 Introduction

In this appendix we investigate the asymptotic properties of the stratified treatment effect estimator, $\hat{\beta}^s$, assuming that the propensity score is estimated from the data using a correctly specified logistic regression model. We begin, in Section B.2, by using standard asymptotic theory to demonstrate that $\hat{\beta}^s$ is a consistent estimator of β_o^s when the propensity score is estimated [1]. We then, in Section B.3, show that $\hat{\beta}^s$ is asymptotically normally distributed. We finish by determining its asymptotic variance using M-estimation theory, in Section B.4. This turns out to be distinct from the asymptotic variance of the stratified treatment effect estimator obtained when the propensity score is a known function of the observed covariates.

This appendix uses the notation given in Section 3.1.1. We now write the propensity score as $p(\mathbf{X}; \boldsymbol{\alpha}_o)$ and $p(\mathbf{X}; \hat{\boldsymbol{\alpha}})$ when the propensity score parameters are known and estimated, respectively, rather than the notation that we used before, $\hat{p}(\mathbf{X})$ and $p_o(\mathbf{X})$, in order to emphasize that the propensity score is now an estimator depending on the unknown parameters $\boldsymbol{\alpha}$. In order to ease our way into the proof, we give a brief review of the estimator, $\hat{\beta}^s$, and the estimating equations used to obtain it.

B.1.1 Estimating the stratified treatment effect

In the notation of Section 3.1.1, where, for subject i, Y_i and Z_i denote the outcome and treatment, $\hat{S_{si}} = 1_{[\hat{q}_{(s-1)} \leq p(\mathbf{X}; \hat{\boldsymbol{\alpha}}) < \hat{q}_s]}$ is an indicator for the s^{th} sample stratum, r_s is the fraction of the sample in the s^{th} stratum, and $\hat{d}_s$ is the fraction of the sample who

[1] We show that $\hat{\beta}^s$ is consistent for the population parameter β_o^s. Since this is usually different from the population average causal treatment effect, β_o, this means that $\hat{\beta}^s$ will usually not be consistent for β_o.

are treated and in the s^{th} sample stratum, the stratified treatment effect estimator is

$$\hat{\beta}^s = \frac{1}{n} \sum_{s=1}^{K} r_s \sum_{i=1}^{n} \left(\frac{Y_i \, Z_i \, \hat{S}_{si}}{\hat{d}_s} - \frac{Y_i \, (1 - Z_i) \, \hat{S}_{si}}{r_s - \hat{d}_s} \right).$$

We assume that the propensity score is related to the observed covariates as follows,

$$\ln \left\{ \frac{p(\mathbf{X}; \boldsymbol{\alpha}_o)}{1 - p(\mathbf{X}; \boldsymbol{\alpha}_o)} \right\} = \boldsymbol{\alpha}_o^T \mathbf{X}. \tag{B.1}$$

As discussed in Chapter 3, we obtain $\hat{\beta}^s$ as a component of the vector solution to the set of estimating equations

$$\sum_{i=1}^{n} \boldsymbol{\psi}(Y_i, Z_i, \mathbf{X}_i \, ; \beta^s, \mathbf{d}, \mathbf{q}) = 0 \tag{B.2}$$

The vector solution is $\hat{\boldsymbol{\theta}}$ defined by $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T, \hat{\boldsymbol{\alpha}}^T)$, where the population values of these parameters are: $\boldsymbol{\alpha}_o$, the population propensity score parameters, $\mathbf{q}_o$, the population strata boundaries, $\mathbf{d}_o$, the population probabilities of being both treated and in each stratum, and β_o^s, the 'true' stratified treatment effect.

The estimating equations (B.2) are defined by $\boldsymbol{\psi}^{(2K+m)\times 1}$ where $\boldsymbol{\psi}^T = (\psi_1, \boldsymbol{\psi}_2^T, \boldsymbol{\psi}_3^T, \boldsymbol{\psi}_4^T)$, with

$$\psi_1^{1\times 1} \, (Y_i, Z_i, \mathbf{X}_i \, ; \beta^s, \mathbf{d}, \mathbf{q}) = \left(\sum_{s=1}^{K} r_s \left\{ \frac{Y_i \, Z_i \, S_{si}}{d_s} - \frac{Y_i \, (1 - Z_i) \, S_{si}}{r_s - d_s} \right\} - \beta^s \right)$$

$$\boldsymbol{\psi}_2^{K\times 1} \, (Z_i, \mathbf{X}_i \, ; \mathbf{d}, \mathbf{q}) = \begin{pmatrix} Z_i \, S_{1i} - d_1 \\ \vdots \\ Z_i \, S_{Ki} - d_K \end{pmatrix}$$

$$\boldsymbol{\psi}_3^{(K-1)\times 1} \, (\mathbf{X}_i \, ; \mathbf{q}) = \begin{pmatrix} S_{1i} - r_1 \\ \vdots \\ S_{(K-1)i} - r_{(K-1)} \end{pmatrix}$$

$$\psi_4^{m\times 1}(Z_i, \mathbf{X}_i; \boldsymbol{\alpha}) = \begin{pmatrix} (Z_i - \frac{\exp(\boldsymbol{\alpha}^T \mathbf{X}_i)}{1+\exp(\boldsymbol{\alpha}^T \mathbf{X}_i)}) X_{1i} \\ \vdots \\ (Z_i - \frac{\exp(\boldsymbol{\alpha}^T \mathbf{X}_i)}{1+\exp(\boldsymbol{\alpha}^T \mathbf{X}_i)}) X_{mi} \end{pmatrix},$$

where the stratum indicator, $S_{si} = 1_{[q_{(s-1)} \leq p(\mathbf{X};\boldsymbol{\alpha}) < q_s]}$ is a function of both the unknown propensity score parameters and the unknown strata boundaries. We also define $q_0 = 0$ and $q_K = 1$. We now use this representation of the estimation process to ascertain the theoretical properties of the stratified treatment effect estimator, β^s.

B.2 Consistency

We now show that each component of the estimator $\hat{\boldsymbol{\theta}}$ is consistent, where $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T, \hat{\boldsymbol{\alpha}}^T)$. We begin by establishing the consistency of the estimated propensity score parameters, $\hat{\boldsymbol{\alpha}}$. We then consider the consistency of the estimated strata boundaries, $\hat{\mathbf{q}}$, treating the propensity score parameters as nuisance parameters. We then consider the consistency of the estimated probabilities of being treated and in each stratum, $\hat{\mathbf{d}}$, treating both the strata boundaries and the propensity score parameters as nuisance parameters. Given the consistency of $\hat{\boldsymbol{\alpha}}, \hat{\mathbf{q}}$ and $\hat{\mathbf{d}}$, we then demonstrate the consistency of the stratified treatment effect estimator, β^s.

B.2.1 Consistency of the estimated propensity score parameters

The propensity score parameters, $\boldsymbol{\alpha}$, are estimated using a maximum likelihood logistic regression model. These parameters are known to be consistent and globally identifiable [17], provided that the model (B.1) is correctly specified.

B.2.2 Consistency of the estimated strata boundaries

We demonstrate the consistency of the estimate of the first strata boundary, $\hat{q}_1$. This is the r_1^{th} quantile of the sample distribution of the estimated propensity score. The population parameter we wish to estimate, q_{1o}, is the r_1^{th} quantile of the population distribution of the propensity score. In order to demonstrate consistency of this estimated strata boundary, when the propensity score parameters are consistently estimated, we refer to Theorem A.1. In order to apply this theorem we need to find a (criteria) function of q_1 and the data that is maximised by the sample quantile, $\hat{q}_1$.

The r_1^{th} quantile of the sample distribution of the estimated propensity score, $\hat{q}_1$. can be expressed as the value of q_1 that maximises the equation

$$M_n(q_1;\hat{\boldsymbol{\alpha}}) \equiv -\sum_{i=1}^{n}\left\{(1-r_1)(q_1 - p(\mathbf{X}_i;\hat{\boldsymbol{\alpha}}))1_{[p(\mathbf{X}_i;\hat{\boldsymbol{\alpha}})<q_1]} + r_1(p(\mathbf{X}_i;\hat{\boldsymbol{\alpha}}) - q_1)1_{[p(\mathbf{X}_i;\hat{\boldsymbol{\alpha}})>q_1]}\right\}$$

We now show that the conditions of Theorem A.1 are satisfied. The function $M_n(q_1;\boldsymbol{\alpha})$ is concave in q_1 and so condition *(i)* of the theorem is satisfied. We now need to show that the value of q_1 that maximises the function $M_n(q_1;\boldsymbol{\alpha}_o)$ is globally identifiable and equal to the r_1^{th} quantile of the population distribution of the propensity score, q_{1o}. By the law of large numbers, as $n \to \infty$, for any q_1 and $\boldsymbol{\alpha}$,

$$\begin{aligned} M_n(q_1;\boldsymbol{\alpha}) &\xrightarrow{p} M(q_1;\boldsymbol{\alpha}) \\ &= \{\mathbb{E}\left[p\left(\mathbf{X};\boldsymbol{\alpha}\right)1_{[p\left(\mathbf{X};\boldsymbol{\alpha}\right)<q_1]}\right] - q_1\mathbb{P}(p\left(\mathbf{X};\boldsymbol{\alpha}\right) < q_1) - r_1\mathbb{E}\left[p\left(\mathbf{X};\boldsymbol{\alpha}\right)\right] + q_1 r_1\}. \end{aligned}$$

Given $\boldsymbol{\alpha}$, since $M(q_1;\boldsymbol{\alpha})$ is a convex function of q_1, this is maximised at a point where $\frac{\partial M(q_1;\boldsymbol{\alpha})}{\partial q_1} = 0$. Differentiating $M_n(q_1;\boldsymbol{\alpha})$ using the product rule and appealing to the fundamental theorem of calculus,

$$\frac{\partial M(q_1;\boldsymbol{\alpha})}{\partial q_1} = q_1\ f_p(q_1;\boldsymbol{\alpha}) - \mathbb{P}(p\left(\mathbf{X};\boldsymbol{\alpha}\right) < q_1) - q_1\ f_p(q_1;\boldsymbol{\alpha}) + r_1,$$

provided that the probability density function of the propensity score, $f_p(\cdot)$, is continuous. Setting $\frac{\partial M(q_1;\boldsymbol{\alpha}_o)}{\partial q_1} = 0$ shows that $M(q_1;\boldsymbol{\alpha}_o)$ is maximised by the value of q_1 such that $\mathbb{P}(p\left(\mathbf{X};\boldsymbol{\alpha}_o\right) < q_1) = r_1$. This value of q_1 is exactly the parameter we wish to estimate, the r_1^{th} quantile of the population distribution of the propensity score, q_{1o}. As before, provided that the cumulative density function of the propensity score is strictly monotone and continuous at the population strata boundary, q_{1o}, this point is unique and therefore globally identifiable. So condition *(ii)* of Theorem A.1 is satisfied.

We now merely have to show that condition *(iii*)* of Theorem A.1 is satisfied in order to prove that the estimated strata boundary, $\hat{q}_1$, is consistent when the propensity score parameters, $\boldsymbol{\alpha}$, are estimated. To do this, we need to show that each of the parameters α_k, for $k = 1, ..., m$, satisfies the Lipschitz condition with respect to the criteria function $M(q_1;\boldsymbol{\alpha})$. We do this by showing that the first derivative of $M(q_1;\boldsymbol{\alpha})$ with respect to α_k, is bounded, since this guarantees condition *(iii*)* [6]. Inspection

of the above formula for $M(q_1;\boldsymbol{\alpha})$ shows that its first derivative with respect to α_k will be bounded provided that the following three functions are bounded,

$$\frac{\partial\, \mathbb{E}\left[\, p\left(\mathbf{X};\boldsymbol{\alpha}\right) 1_{\left[p\left(\mathbf{X};\boldsymbol{\alpha}\right)<q_1\right]}\,\right]}{\partial \alpha_k}, \qquad \frac{\partial\, \mathbb{P}(p\left(\mathbf{X};\boldsymbol{\alpha}\right)<q_1)}{\partial \alpha_k} \qquad \text{and} \qquad \frac{\partial\, \mathbb{E}\left[\, p\left(\mathbf{X};\boldsymbol{\alpha}\right)\right]}{\partial \alpha_k}.$$

We can write the first of these three functions in integral form as follows,

$$\frac{\partial}{\partial \alpha_k}\{\mathbb{E}\left[\, p\left(\mathbf{X};\boldsymbol{\alpha}\right) 1_{\left[p\left(\mathbf{X};\boldsymbol{\alpha}\right)<q_1\right]}\,\right]\} = \frac{\partial}{\partial \alpha_k}\left\{\int_0^{q_1} p\; f_p(p;\boldsymbol{\alpha})\; dp\right\}.$$

Provided that the probability density function of the propensity score is continuous in $\boldsymbol{\alpha}$, in other words, continuous in the propensity score, we can exchange the order of integration and differentiation, giving

$$\frac{\partial}{\partial \alpha_k}\{\mathbb{E}\left[\, p\left(\mathbf{X};\boldsymbol{\alpha}\right) 1_{\left[p\left(\mathbf{X};\boldsymbol{\alpha}\right)<q_1\right]}\,\right]\} = \int_0^{q_1} p\; \frac{\partial}{\partial \alpha_k}\{f_p(p;\boldsymbol{\alpha})\; dp\} \leq\; q_1\; \sup_{p\in[0,q_1]}\left\{\frac{\partial f_p(p;\boldsymbol{\alpha})}{\partial \alpha_k}\right\}.$$

So this derivative will be bounded provided that the derivative $\frac{\partial f_p(p;\boldsymbol{\alpha})}{\partial \alpha_k}$ is bounded. The same argument shows that the same conditions guarantee that $\frac{\partial M(q_1;\boldsymbol{\alpha})}{\partial \alpha_k}$ is bounded. Then the first estimated strata boundary, $\hat{q}_1$, is a consistent estimator of the first population boundary, q_{1o}. In the same way, we can show that all the estimated strata boundaries, $\hat{\mathbf{q}}$, are consistent estimators of the population strata boundaries, $\mathbf{q}_o$, under consistent estimation of the propensity score parameters.

B.2.3 Consistency of the estimated probabilities of being treated and in each stratum

We now show that the estimated probabilities of being treated and in each stratum, $\hat{\mathbf{d}}$, are consistent estimators of the analogous population probabilities, $\mathbf{d}_o$, when the strata boundaries and the propensity score parameters are consistently estimated. In particular, we consider the consistency of $\hat{d}_1$, the estimated probability of being treated and in the first stratum. We again appeal to Theorem A.1 in order to prove this. We have seen that the parameters $\mathbf{d}_o, \mathbf{q}_o$, and $\boldsymbol{\alpha}_o$ are globally identifiable, so condition *(ii)* of Theorem A.1 is satisfied. We saw in Appendix A.2.2 that when the propensity score is known, the estimator $\hat{d}_1$ can be obtained by finding the value of d_1 that maximises a particular criteria function. When the propensity score is estimated, we view the criteria function as a function also of the estimated propensity score parameters, $\hat{\boldsymbol{\alpha}}$, and then we can obtain $\hat{d}_1$ from the same criteria function,

$M_n(d_1; \hat{q}_1, \hat{\boldsymbol{\alpha}})$. This function is concave in d_1 and $M_n(d_1; q_1, \boldsymbol{\alpha}) \xrightarrow{p} M(d_1; q_1, \boldsymbol{\alpha})$ as $n \to \infty$, where

$$M(d_1; q_1, \boldsymbol{\alpha}) = (1 - 2d_1) \int_0^{q_1} p \; f_p(p; \boldsymbol{\alpha}) \; dp + d_1^2.$$

In order to show that $\hat{d}_1$ is consistent when the propensity score parameters are estimated we merely need to show that the Lipschitz condition holds for each α_k with respect to the function $M(d_1; q_1, \boldsymbol{\alpha})$. It is sufficient to show that $\frac{\partial M(d_1; q_1, \boldsymbol{\alpha})}{\partial \alpha_k}$ is bounded for each $k = 1, ..., m$. We can see immediately that this will be true provided that the the probability density function of the propensity score is continuous and that the derivatives $\frac{\partial f_p(p; \boldsymbol{\alpha})}{\partial \alpha_k}$ exist and are bounded. Then the estimated probabilities of being treated and in each stratum, $\hat{\mathbf{d}}$, are consistent estimators of $\mathbf{d}_o$ under consistent estimation of the strata boundaries and propensity score parameters.

B.2.4 Consistency of the stratified treatment effect estimator

We now show that the stratified treatment effect estimator, $\hat{\beta}^s$, is a consistent estimator of β_o^s when the propensity score parameters are estimated using a correctly specified logistic regression model. We saw in Appendix A.2.3 that $\hat{\beta}^s$ can be obtained by finding the value of β^s that maximises a criteria function $M_n(\beta^s; \hat{\mathbf{q}}, \hat{\mathbf{d}}, \hat{\boldsymbol{\alpha}})$ that is concave in β^s, where

$$M_n(\beta^s; \mathbf{q}, \mathbf{d}, \boldsymbol{\alpha}) \xrightarrow{p} M(\beta^s; \mathbf{q}, \mathbf{d}, \boldsymbol{\alpha})$$

with

$$M(\beta^s; \mathbf{q}, \mathbf{d}, \boldsymbol{\alpha}) = -(\beta^s)^2 + 2\beta^s \sum_{s=1}^{K} r_s \left\{ \frac{\mathbb{E}[Y \, Z \, S_s]}{d_s} - \frac{\mathbb{E}[Y \, (1 - Z) \, S_s]}{r_s - d_s} \right\}$$
$$+ \sum_{s=1}^{K} r_s^2 \left\{ \frac{\mathbb{E}[Y^2 \, Z \, S_s]}{d_s^2} + \frac{\mathbb{E}[Y^2 \, (1 - Z) \, S_s]}{(r_s - d_s)^2} \right\}.$$

We have already seen that each of the unknown parameters is globally identifiable. Therefore, conditions *(i)* and *(ii)* of Theorem A.1 are satisfied. We now need to show that the Lipschitz condition is satisfied for each propensity score parameter α_k, for $k = 1, ..., m$. In order to do this we show that the first derivative of $M(\beta^s; \mathbf{q}, \mathbf{d}, \boldsymbol{\alpha})$ with respect to α_k, for $k = 1, .., m$, is bounded. We can see immediately that this will

be true if the following functions are bounded,

$$\frac{\partial\, \mathbb{E}\,[Y\, Z\, S_s]}{\partial \alpha_k}, \qquad \frac{\partial\, \mathbb{E}\,[Y\,(1-Z)\, S_s]}{\partial \alpha_k}$$
$$\frac{\partial\, \mathbb{E}\,[Y^2\, Z\, S_s]}{\partial \alpha_k}, \qquad \frac{\partial\, \mathbb{E}\,[Y^2\,(1-Z)\, S_s]}{\partial \alpha_k}$$

By writing each of these expectations as an integral over the propensity score values in a stratum, interchanging the order of integration and differentiation and appealing to the fundamental theorem of calculus, we see that the four functions above will be bounded provided that

- the probability density function of the propensity score is continuous and bounded everywhere;
- the derivatives $\frac{\partial f_p(\cdot)}{\partial \alpha_k}$ exist and are bounded everywhere, for $k = 1, ..., m$;
- the following expectations are continuous in p and bounded everywhere, for $t = 0, 1$,

$$\mathbb{E}\,[Y \mid Z = t,\, p\,(\mathbf{X}) = p\,], \qquad \mathbb{E}\,[Y^2 \mid Z = t,\, p\,(\mathbf{X}) = p\,];$$

- the following derivatives exist and are bounded, for $t = 0, 1$ and $k = 1, ..., m$,

$$\frac{\partial}{\partial \alpha_k}\{\mathbb{E}\,[Y \mid Z = t,\, p\,(\mathbf{X}) = p\,]\}, \qquad \frac{\partial}{\partial \alpha_k}\{\mathbb{E}\,[Y^2 \mid Z = t,\, p\,(\mathbf{X}) = p\,]\}.$$

Then all the conditions have been satisfied and so the stratified treatment effect estimator is a consistent estimator of β_o^s. The implications of the conditions attached to this consistency are discussed in the main text (Section 3.3).

B.3 Asymptotic normality

We now consider the asymptotic sampling distribution of the estimator $\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T, \hat{\boldsymbol{\alpha}}^T)$. We refer to Theorem A.2 which states conditions under which an estimator, $\hat{\boldsymbol{\theta}}$, obtained by solving the estimating equation $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{X}_i; \boldsymbol{\theta}) = 0$, is asymptotically normal. Our estimating equation is defined by $\boldsymbol{\psi}^T = (\psi_1, \boldsymbol{\psi}_2^T, \boldsymbol{\psi}_3^T, \boldsymbol{\psi}_4^T)$, where $\psi_1, \boldsymbol{\psi}_2, \boldsymbol{\psi}_3$ and $\boldsymbol{\psi}_4$ are defined in Section B.1. We now ensure that all conditions of Theorem A.2 are satisfied. We have already demonstrated consistency of $\hat{\boldsymbol{\theta}}$, and that the maps $\mathbf{W} \rightarrow \boldsymbol{\psi}_j(\mathbf{W}; \boldsymbol{\theta})$ are Donsker, for $j = 1, 2, 3$ (see Section A.3). We

also need to show that the map $\mathbf{W} \rightarrow \boldsymbol{\psi}_4(\mathbf{W};\boldsymbol{\theta})$ is Donsker, where $\mathbf{W} = (Y, Z, \mathbf{X})$ represents the data. We then show that the maps $\boldsymbol{\theta} \rightarrow \boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta})$ are continuous in probability. The final condition of the theorem is that the derivative matrix, $\frac{\partial}{\partial \boldsymbol{\theta}^T} \{\mathbb{E}[\,\boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta})\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$, exists and is non-singular. This will be demonstrated in Section B.4.

The map $\mathbf{W} \rightarrow \boldsymbol{\psi}_4(\mathbf{W};\boldsymbol{\theta})$ is defined as $\boldsymbol{\alpha} \rightarrow \left(Z - \frac{\exp(\boldsymbol{\alpha}^T\mathbf{X})}{1+\exp(\boldsymbol{\alpha}^T\mathbf{X})}\right)\mathbf{X}$. Example 19.7 of van der Vaart [108, p.271] states that a function of $\boldsymbol{\theta}$ satisfying the Lipschitz condition with respect to $\boldsymbol{\theta}$ is Donsker. The first derivative of the map $\mathbf{W} \rightarrow \boldsymbol{\psi}_4(\mathbf{W})$ with respect to α_k is bounded for $k = 1, ..., m$, and so the Lipschitz condition is satisfied. Therefore, this map is Donsker. Note that the same condition guarantees that the map $\boldsymbol{\theta} \rightarrow \boldsymbol{\psi}_4(\mathbf{W};\boldsymbol{\theta})$ is continuous in probability.

We now show that the map $\boldsymbol{\theta} \rightarrow \boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta})$ is continuous in probability. If we define $\boldsymbol{\theta}^T = (\beta^s, \mathbf{q}^T, \mathbf{d}^T, \boldsymbol{\alpha}^T)$ and $\boldsymbol{\theta}'^T = ((\beta^s)', \mathbf{q}'^T, \mathbf{d}'^T, \boldsymbol{\alpha}'^T)$, then we need to show that

$$\mathbb{E}[\,|\boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta}) - \boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta}')|\,] \rightarrow 0$$

as $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}'$. Using the definition of $\boldsymbol{\psi}_3$ given in Section B.1,

$$\mathbb{E}[\,|\boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta}) - \boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta}')|\,] = \mathbb{E}[\,|1_{[q_{(j-1)} \leq (\mathbf{X};\boldsymbol{\alpha}) < q_j]} - 1_{[q'_{(j-1)} \leq p(\mathbf{X};\boldsymbol{\alpha}') < q'_j]}|\,].$$

Using the triangle inequality,

$$\mathbb{E}[\,|\boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta}) - \boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta}')|\,] \;\leq\; e_1 + e_2 + e_3,$$

where

$$\begin{aligned} e_1 &= \mathbb{E}[\,|1_{[q_{(j-1)} \leq p(\mathbf{X};\boldsymbol{\alpha}) < q_j]} - 1_{[q_{(j-1)} \leq p(\mathbf{X};\boldsymbol{\alpha}) < q'_j]}|\,] \\ e_2 &= \mathbb{E}[\,|1_{[q_{(j-1)} \leq p(\mathbf{X};\boldsymbol{\alpha}) < q'_j]} - 1_{[q'_{(j-1)} \leq p(\mathbf{X};\boldsymbol{\alpha}) < q'_j]}|\,] \\ e_3 &= \mathbb{E}[\,|1_{[q'_{(j-1)} \leq p(\mathbf{X};\boldsymbol{\alpha}) < q'_j]} - 1_{[q'_{(j-1)} \leq p(\mathbf{X};\boldsymbol{\alpha}') < q'_j]}|\,]. \end{aligned}$$

We saw in Appendix A.3 that e_1 and e_2 tend to zero as $\mathbf{q} \rightarrow \mathbf{q}'$, provided that the probability density function of the propensity score is bounded. We can write e_3 in

integral form as

$$e_3 = \int_{[q'_{(j-1)}, q'_j]} \{ f_p(p;\boldsymbol{\alpha}) + f_p(p;\boldsymbol{\alpha}') \}\, dp \;\rightarrow\; 0,$$

as $\boldsymbol{\alpha} \rightarrow \boldsymbol{\alpha}'$, provided that the probability density function of the propensity score, $f_p(\,\cdot\,)$, is continuous in $\boldsymbol{\alpha}$. Then $\boldsymbol{\theta} \rightarrow \boldsymbol{\psi}_{3j}(\mathbf{W};\boldsymbol{\theta})$ is continuous in probability.

In the same way we can see that the maps $\boldsymbol{\theta} \rightarrow \boldsymbol{\psi}_{2k}(\mathbf{W};\boldsymbol{\theta})$ and $\boldsymbol{\theta} \rightarrow \psi_1(\mathbf{W};\boldsymbol{\theta})$ are also continuous in probability, for $k = 1, ..., K$, provided that the probability density function of the propensity score, $f_p(\,\cdot\,)$, is continuous and bounded.

The final requirement of Theorem A.2 is that the derivative matrix $\frac{\partial}{\partial \boldsymbol{\theta}^T} \{\boldsymbol{\psi}(\mathbf{W};\boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$ should exist and be non-singular. This matrix is calculated in Section B.4, and conditions under which it exists and is invertible are stated. Provided that this inverse exists and the conditions we have found are satisfied, then the conditions of Theorem A.2 are satisfied and $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed.

B.4 Asymptotic variance

We have now established conditions under which the estimator $\hat{\boldsymbol{\theta}}$, defined by $\hat{\boldsymbol{\theta}}^T = (\hat{\beta}^s, \hat{\mathbf{d}}^T, \hat{\mathbf{q}}^T, \hat{\boldsymbol{\alpha}})$, is consistent and asymptotically normally distributed. We now use its asymptotic distribution to calculate the asymptotic variance of the stratified treatment effect estimator, $\hat{\beta}^s$, assuming that the propensity score is estimated using a correctly specified logistic regression model. We denote this variance by $\mathbb{V}_e[\,\hat{\beta}^s\,]$, where the 'e' refers to the estimation of the propensity score.

B.4.1 M-estimation theory

The general M-estimation theory outlined in the main text (Section 3.1.2) shows that if we let

$$\begin{aligned} A &= \mathbb{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}^T} \{\boldsymbol{\psi}(Y, Z, \mathbf{X};\, \boldsymbol{\theta})\} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \right] \\ B &= \mathbb{E}\left[\, \boldsymbol{\psi}(Y, Z, \mathbf{X};\, \boldsymbol{\theta}_o)\; \boldsymbol{\psi}^T(Y, Z, \mathbf{X};\, \boldsymbol{\theta}_o)\, \right], \end{aligned} \tag{B.3}$$

then the large-sample covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by

$$\text{Cov}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\,\mathbf{A^{-1}B A^{-T}}. \tag{B.4}$$

We furthermore stated that when the function $\boldsymbol{\psi}(Y, Z, \mathbf{X};\, \boldsymbol{\theta})$ is not differentiable with respect to $\boldsymbol{\theta}$, as in our example, the order of differentiation and expectation can be exchanged and so

$$A \;= -\frac{\partial}{\partial \boldsymbol{\theta}^T}\{\mathbb{E}\,[\,\boldsymbol{\psi}(Y, Z, \mathbf{X};\, \boldsymbol{\theta})\,]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}. \tag{B.5}$$

However, this is only a valid approach under certain regularity conditions. Although we do not list these regularity conditions, the application of Theorem A.2 shows that it is valid in our problem. Therefore, when a component of $\boldsymbol{\psi}(Y, Z, \mathbf{X};\, \boldsymbol{\theta})$ is not differentiable with respect to $\boldsymbol{\theta}$ we use (B.5). For components of $\boldsymbol{\psi}(Y, Z, \mathbf{X};\, \boldsymbol{\theta})$ that are differentiable with respect to $\boldsymbol{\theta}$, however, since the two versions of A are then identical, we usually use (B.3), since this is often more convenient.

Allowing for definition (B.5) to be used if necessary, (B.4) is then equal to the variance given at the end of Theorem A.2, and the large-sample variance of the stratified treatment effect estimator, $\hat{\beta}^s$ is

$$\mathbb{V}_e[\,\hat{\beta}^s\,] = \mathbb{V}\,[\,(\hat{\boldsymbol{\theta}})_1\,] = \frac{1}{n}\,(\mathbf{A^{-1}B A^{-T}})_{11}.$$

We have partitioned the parameter $\boldsymbol{\theta}$ into four components, $\boldsymbol{\theta}^T = (\beta^s, \mathbf{d}^T, \mathbf{q}^T, \boldsymbol{\alpha}^T)$. The matrices $A^{(2K+m)\times(2K+m)}$ and $B^{(2K+m)\times(2K+m)}$ can be partitioned in the same way. Then

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix},$$

where for $j = 1, 2, 3, 4$,

$$a_{j1} = -\frac{\partial}{\partial \beta^s}\,\{\mathbb{E}\,[\,\boldsymbol{\psi}_j\,]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \qquad a_{j2} = -\frac{\partial}{\partial \mathbf{d}^T}\,\{\mathbb{E}\,[\,\boldsymbol{\psi}_j\,]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$

$$a_{j3} = -\frac{\partial}{\partial \mathbf{q}^T}\,\{\mathbb{E}\,[\,\boldsymbol{\psi}_j\,]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \qquad a_{j4} = -\frac{\partial}{\partial \boldsymbol{\alpha}^T}\,\{\mathbb{E}\,[\,\boldsymbol{\psi}_j\,]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

These sub-matrices of A have the following dimensions:

$$Dim = \begin{pmatrix} 1 \times 1 & 1 \times K & 1 \times (K-1) & 1 \times m \\ K \times 1 & K \times K & K \times (K-1) & K \times m \\ (K-1) \times 1 & (K-1) \times K & (K-1) \times (K-1) & (K-1) \times m \\ m \times 1 & m \times K & m \times (K-1) & m \times m \end{pmatrix}, \quad \text{(B.6)}$$

where m is the number of covariates included in the logistic regression model used to estimate the propensity score. We have already defined the sub-matrices a_{ij} for $i, j = 1, 2, 3$, in Appendix A.4.2. Since the function $\boldsymbol{\psi}_4$ does not contain β^s, $\mathbf{d}$ or $\mathbf{q}$, differentiating with respect these parameters results in zero. Therefore, a_{41}, a_{42} and a_{43} are zero matrices. Then A is

$$A = \begin{pmatrix} 1 & a_{12} & a_{13} & a_{14} \\ 0 & I & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{pmatrix},$$

which has inverse

$$A^{-1} = \begin{pmatrix} 1 & -a_{12} & (a_{12}a_{23} - a_{13})a_{33}^{-1} & a^* \\ 0 & I & -a_{23}a_{33}^{-1} & (a_{23}a_{33}^{-1}a_{34} - a_{24})a_{44}^{-1} \\ 0 & 0 & a_{33}^{-1} & -a_{33}^{-1}a_{34}a_{44}^{-1} \\ 0 & 0 & 0 & a_{44}^{-1} \end{pmatrix},$$

where $a^* = (a_{12}a_{24} - a_{14})a_{44}^{-1} - (a_{12}a_{23} - a_{13})a_{33}^{-1}a_{34}a_{44}^{-1}$. The conditions under which the sub-matrices a_{12}, a_{13}, a_{23}, and a_{33}^{-1} exist have already been investigated and are stated in Appendix A.4.2. Conditions under which the sub-matrices a_{14}, a_{24}, a_{34} and a_{44}^{-1} exist are investigated later in this section. When these sub-matrices exist, so does the inverse A^{-1}. The existence of this inverse is one of the conditions for the asymptotic normality of $\hat{\boldsymbol{\theta}}$.

Partitioning the matrix B in a similar fashion gives

$$B = \begin{pmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \\ b_{41} & b_{42} & b_{43} & b_{44} \end{pmatrix},$$

where $b_{jk} = \mathbb{E}[\boldsymbol{\psi}_j(\boldsymbol{\theta}_o)\ \boldsymbol{\psi}_k^T(\boldsymbol{\theta}_o)]$, for $j, k = 1, 2, 3, 4$. The dimensions of these sub-matrices are also given by (B.6).

We have seen that $n\,\mathbb{V}_e[\,\hat{\beta}^s\,] = (A^{-1}BA^{-T})_{11}$. Multiplying out $A^{-1}BA^{-T}$ and taking the $(1,1)^{th}$ component, remembering that B is symmetric, gives

$$\begin{aligned} n\,\mathbb{V}_e[\,\hat{\beta}^s\,] = {} & b_{11} - 2b_{12}a_{12}^T + a_{12}b_{22}a_{12}^T \\ & + (a_{12}a_{23} - a_{13})a_{33}^{-1}\{2b_{31} - 2b_{32}a_{12}^T + b_{33}a_{33}^{-T}(a_{12}a_{23} - a_{13})^T\} \\ & + 2(b_{14}a^{*T} - a_{12}b_{24}a^{*T} + (a_{12}a_{23} - a_{13})a_{33}^{-1}b_{34}a^{*T}) + a^*b_{44}a^{*T}. \end{aligned} \tag{B.7}$$

We can see that the first two lines of the equation above are exactly the formula for $n\,\mathbb{V}_e[\,\hat{\beta}^s\,]$ that we had previously when the propensity score was known (A.13). The additional terms are as follows,

$$2(b_{14} - a_{12}b_{24} + (a_{12}a_{23} - a_{13})a_{33}^{-1}b_{34})a^{*T} + a^*\ b_{44}\ a^{*T}.$$

These terms represent the additional variance of the stratified treatment effect estimator due to the estimation of the propensity score. The sub-matrices contained in this formula that were not calculated in the previous appendix are calculated in Sections B.4.2 and B.4.3. These sub-matrices are substituted into the variance equation (B.7) and the variance $n\,\mathbb{V}_e[\,\hat{\beta}^s\,]$ is then calculated by direct matrix multiplication.

B.4.2 The matrix A

We now calculate the sub-matrices of A that have not been calculated previously, that is a_{14}, a_{24}, a_{34} and a_{44}^{-1}. We assume that the subjects are sampled independently from the population and so we drop the subject subscript from the following calculations in order to simplify the presentation. In particular, note that then X_i refers to the i^{th} covariate and not the value of covariate X taken by the i^{th} subject.

The sub-matrix $\mathbf{a_{14}}$ Since the function ψ_1 is not differentiable with respect to the propensity score parameters, α_k, we interchange the order of differentiation and expectation and define, for $k = 1, ..., m$,

$$\begin{aligned} (a_{14})_k = {} & -\frac{\partial}{\partial\alpha_k}\ \{\mathbb{E}[\,\psi_1\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \\ & -\frac{\partial}{\partial\alpha_k}\ \mathbb{E}\left[Y\,Z\sum_{s=1}^{K}\frac{r_s\ S_s}{d_s} - Y\,(1-Z)\sum_{s=1}^{k}\frac{r_s\ S_s}{r_s - d_s} - \beta^s \right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}. \end{aligned}$$

Using the equality $\mathbb{E}[Y\,Z\,S_s] = \mathbb{E}[Y \mid Z=1,\, S_s=1]\,\mathbb{E}[ZS_s]$, we get

$$(a_{14})_k = -\sum_{s=1}^{K} r_s \frac{\partial}{\partial \alpha_k}\left\{\frac{\mathbb{E}[Y \mid Z=1,\, S_s=1]\mathbb{E}[Z\,S_s]}{d_s}\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$+\sum_{s=1}^{K} r_s \frac{\partial}{\partial \alpha_k}\left\{\frac{\mathbb{E}[Y \mid Z=0,\, S_s=1]\mathbb{E}[(1-Z)\,S_s]}{r_s-d_s}\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

Differentiating this by the product rule and using the equality $\mathbb{E}[\psi(\boldsymbol{\theta}_o)] = 0$, so $\mathbb{E}[Z\,S_{so}] = d_{so}$, we have, for $k = 1, ..., m$,

$$(a_{14})_k = -\sum_{s=1}^{K} \frac{r_s}{d_{so}}\,\mathbb{E}[Y \mid Z=1,\, S_{so}=1]\,\frac{\partial}{\partial \alpha_k}\,\{\mathbb{E}[ZS_s]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$+\sum_{s=1}^{K} \frac{r_s}{r_s-d_{so}}\,\mathbb{E}[Y \mid Z=0, S_{so}=1]\,\frac{\partial}{\partial \alpha_k}\,\{\mathbb{E}[(1-Z)S_s=1]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$-\sum_{s=1}^{K} r_s \frac{\partial}{\partial \alpha_k}\{\mathbb{E}[Y \mid Z=1,\, S_s=1] - \mathbb{E}[Y \mid Z=0,\, S_s=1]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

The sub-matrix $\mathbf{a_{24}}$ Again, we exchange the order of differentiation and expectation and define, for $j = 1, ..., K$ and $k = 1, ..., m$,

$$(a_{24})_{jk} = -\frac{\partial}{\partial \alpha_k}\,\{\mathbb{E}[\psi_{2j}]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -\frac{\partial}{\partial \alpha_k}\,\{\mathbb{E}[ZS_j - d_j]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$= -\frac{\partial}{\partial \alpha_k}\,\{\mathbb{E}[Z\,1_{[q_{(j-1)} \le p(\mathbf{X};\boldsymbol{\alpha}) < q_j]}]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

The sub-matrix $\mathbf{a_{34}}$ For $j = 1, ..., K-1$ and $k = 1, ..., m$,

$$(a_{34})_{jk} = -\frac{\partial}{\partial \alpha_k}\,\{\mathbb{E}[\psi_{3j}]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -\frac{\partial}{\partial \alpha_k}\,\{\mathbb{E}[S_j - r_j]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$= -\frac{\partial}{\partial \alpha_k}\,\{\mathbb{E}[1_{[q_{j-1} \le p(\mathbf{X};\boldsymbol{\alpha}) < q_j]}]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

The sub-matrix $\mathbf{a_{44}}$ For $j = 1, ..., m$ and $k = 1, ..., m$,

$$(a_{44})_{jk} = -\frac{\partial}{\partial \alpha_k}\,\{\mathbb{E}[\psi_{4j}]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$= -\frac{\partial}{\partial \alpha_k}\left\{\mathbb{E}\left[\left(Z - \frac{\exp(\boldsymbol{\alpha}^T\mathbf{X})}{1+\exp(\boldsymbol{\alpha}^T\mathbf{X})}\right)X_j\right]\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$
$$= \mathbb{E}[p(\mathbf{X};\boldsymbol{\alpha}_o)(1-p(\mathbf{X};\boldsymbol{\alpha}_o))\,X_j\,X_k].$$

Conditions under which A^{-1} exists The inverse of the matrix A will exist if all its components exist. Inspection of the sub-matrices calculated in this section shows that this will happen provided that:

- the probabilities d_s are not equal to 0 or r_s for $s = 1, ..., K$;
- the probability density function of the propensity score is non-zero at each of the population strata boundaries;
- the following derivatives exist, for $t = 0, 1, s = 1, ..., K$ and $j = 1, ..., K-1$:

$$\frac{\partial\, \mathbb{E}\,[\,Y \mid Z = t,\, S_s = 1\,]}{\partial q_j}, \qquad \frac{\partial\, \mathbb{E}\,[\,ZS_s\,]}{\partial q_j}, \qquad \frac{\partial\, \mathbb{E}\,[\,S_s\,]}{\partial q_j}.$$

  By writing these derivatives as integrals over the propensity score, and appealing to the fundamental theorem of calculus, we see that these derivatives exist provided that both the probability density function of the propensity score and $\mathbb{E}\,[\,Y|Z = t,\, p\,(\mathbf{X}; \boldsymbol{\alpha}) = p\,]$ are continuous everywhere in p;

- the following derivatives exist, for $t = 0, 1,\ s = 1, ..., K$ and $k = 1, ..., m$:

$$\frac{\partial\, \mathbb{E}\,[\,Y \mid Z = t,\, S_s = 1\,]}{\partial \alpha_k}, \qquad \frac{\partial\, \mathbb{E}\,[\,ZS_s\,]}{\partial \alpha_k}, \qquad \frac{\partial\, \mathbb{E}\,[\,S_s\,]}{\partial \alpha_k}.$$

  By writing these derivatives as integrals over the propensity score, and appealing to the fundamental theorem of calculus, we see that these derivatives exist provided that $\mathbb{E}\,[\,Y \mid Z = t,\, p\,(\mathbf{X}; \boldsymbol{\alpha}) = p\,]$ is continuous in p and that the following derivatives exist:

$$\frac{\partial\, \mathbb{E}\,[\,Y \mid Z = t,\, p\,(\mathbf{X}; \boldsymbol{\alpha}) = p\,]}{\partial \alpha_k}, \quad \cdots \quad \frac{\partial\, f_p(p; \boldsymbol{\alpha})}{\partial \alpha_k}.$$

B.4.3 The matrix B

In this sub-section, the sub-matrices of B that have not already been calculated, that is b_{14}, b_{24}, b_{34} and b_{44}, are calculated using the formula $b_{jk} = \mathbb{E}\,[\,\boldsymbol{\psi}_j(\boldsymbol{\theta}_o)\ \boldsymbol{\psi}_k^T(\boldsymbol{\theta}_o)\,]$, for $j, k = 1, 2, 3$. The equations $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \boldsymbol{\psi}_3$ and $\boldsymbol{\psi}_4$ are given in Section B.1.

The sub-matrix $\mathbf{b_{14}}$ For $j = 1, ..., m$,

$$
\begin{aligned}
(b_{14})_j &= \mathbb{E}\left[\psi_1\ \psi_{4j}\right] \\
&= \mathbb{E}\left[\sum_{s=1}^{K} r_s \left\{\frac{Y\,Z\,S_{so}}{d_{so}} - \frac{Y\,(1-Z)\,S_{so}}{r_s - d_{so}} - \beta_o^s\right\}\left(Z - \frac{\exp\left(\boldsymbol{\alpha}_o^T\mathbf{X}\right)}{1+\exp\left(\boldsymbol{\alpha}_o^T\mathbf{X}\right)}\right) X_j\right].
\end{aligned}
$$

Remembering that $p\left(\mathbf{X};\boldsymbol{\alpha}_o\right) = \frac{\exp\left(\boldsymbol{\alpha}_o^T\mathbf{X}\right)}{1+\exp\left(\boldsymbol{\alpha}_o^T\mathbf{X}\right)}$ we can write

$$
(b_{14})_j = \sum_{s=1}^{K} r_s\ \mathbb{E}\left[\frac{Y\,Z\,X_j\,S_{so}}{d_{so}} - p\left(\mathbf{X};\boldsymbol{\alpha}_o\right) X_j\left(\frac{Y\,Z\,S_{so}}{d_{so}} - \frac{Y\,(1-Z)\,S_{so}}{r_s - d_{so}}\right)\right].
$$

And using the equality $\mathbb{E}\left[Y\,Z\,S_{so}\right] = \mathbb{E}\left[Y \mid Z=1, S_{so}=1\right] d_{so}$, we have

$$
\begin{aligned}
(b_{14})_j = \sum_{s=1}^{K} r_s \{&\mathbb{E}\left[Y\,X_j \mid Z=1,\ S_{so}=1\right] - \mathbb{E}\left[Y\,p\left(\mathbf{X};\boldsymbol{\alpha}_o\right) X_j \mid Z=1,\ S_{so}=1\right] \\
&-\mathbb{E}\left[Y\,p\left(\mathbf{X};\boldsymbol{\alpha}_o\right) X_j \mid Z=0,\ S_{so}=1\right]\}.
\end{aligned}
$$

Then for $j = 1, ..., m$,

$$
\begin{aligned}
(b_{14})_j = &\sum_{s=1}^{K} r_s\ \mathbb{E}\left[Y\,X_j\,\left(1-p\left(\mathbf{X};\boldsymbol{\alpha}_o\right)\right) \mid Z=1,\ S_{so}=1\right] \\
&+\sum_{s=1}^{K} r_s\ \mathbb{E}\left[Y\,X_j\,p\left(\mathbf{X};\boldsymbol{\alpha}_o\right) \mid Z=0,\ S_{so}=1\right].
\end{aligned}
$$

The sub-matrix $\mathbf{b_{24}}$ For $j = 1, ..., K$ and $k = 1, ..., m$,

$$
\begin{aligned}
(b_{24})_{jk} &= \mathbb{E}\left[\psi_{2j}^T\ \psi_{4k}\right] \\
&= \mathbb{E}\left[\left(Z\,S_{jo} - d_{jo}\right)\left(Z - \frac{\exp\left(\boldsymbol{\alpha}_o^T\mathbf{X}\right)}{1+\exp\left(\boldsymbol{\alpha}_o^T\mathbf{X}\right)}\right) X_k\right] \\
&= \mathbb{E}\left[p\left(\mathbf{X};\boldsymbol{\alpha}_o\right)\left(1-p\left(\mathbf{X};\boldsymbol{\alpha}_o\right)\right)\ X_k\ S_{jo}\right].
\end{aligned}
$$

The sub-matrix $\mathbf{b_{34}}$ For $j = 1, ..., K-1$ and $k = 1, ..., m$,

$$
\begin{aligned}
(b_{34})_{jk} &= \mathbb{E}\left[\psi_{3j}^T\ \psi_{4k}\right] \\
&= \mathbb{E}\left[\left(S_{jo} - r_j\right)\left(Z - \frac{\exp\left(\boldsymbol{\alpha}_o^T\mathbf{X}\right)}{1+\exp\left(\boldsymbol{\alpha}_o^T\mathbf{X}\right)}\right) X_k\right] \\
&= \mathbb{E}\left[Z\,X_k\,S_{jo}\right] - \mathbb{E}\left[p\left(\mathbf{X};\boldsymbol{\alpha}_o\right) X_k\,S_{jo}\right] = 0,
\end{aligned}
$$

where the last line was reached since by conditioning on the observed covariates, $\mathbf{X}$.

The sub-matrix $\mathbf{b_{44}}$ For $j = 1, ..., m$ and $k = 1, ... m$, remembering that the propensity score is defined as $p(\mathbf{X}; \boldsymbol{\alpha}_o) = \frac{\exp(\boldsymbol{\alpha}_o^T \mathbf{X})}{1+\exp(\boldsymbol{\alpha}_o^T \mathbf{X})}$,

$$\begin{aligned}(b_{44})_{jk} &= \mathbb{E}[\,\psi_{4j}^T\,\psi_{4k}\,] \\ &= \mathbb{E}\left[\left(Z - \frac{\exp(\boldsymbol{\alpha}_o^T\mathbf{X})}{1+\exp(\boldsymbol{\alpha}_o^T\mathbf{X})}\right)^2 X_j\,X_k\right] \\ &= \mathbb{E}[\,p(\mathbf{X};\boldsymbol{\alpha}_o)(1-p(\mathbf{X};\boldsymbol{\alpha}_o))\,X_j\,X_k\,],\end{aligned}$$

When a maximum likelihood regression model is used, the matrices A and B are identical. Since the propensity score parameters are estimated using a maximum likelihood logistic regression model, we therefore expect the matrices a_{44} and b_{44} to be equal. Comparing these two matrices shows that this is indeed the case.

B.4.4 Variance of the stratified treatment effect estimator

We have already seen that

$$\begin{aligned}n\,\mathbb{V}_e[\,\hat{\beta}^s\,] &= b_{11} - 2b_{12}a_{12}^T + a_{12}b_{22}a_{12}^T \\ &+ (a_{12}a_{23} - a_{13})a_{33}^{-1}\{2b_{31} - 2b_{32}a_{12}^T + b_{33}a_{33}^{-T}(a_{12}a_{23} - a_{13})^T\} \\ &+ 2(b_{14}a^{*T} - a_{12}b_{24}a^{*T} + (a_{12}a_{23} - a_{13})a_{33}^{-1}b_{34}a^{*T}) + a^*b_{44}a^{*T}.\end{aligned}$$

The sub-matrices involved in this formula have been calculated. Some rather lengthy direct matrix multiplication of the right-hand side of this equation shows that

$$\mathbb{V}_e[\,\hat{\beta}^s\,] = V_1 + V_2 + \frac{2}{n}\,\mathbf{C}\,(n\,\mathrm{Cov}[\,\hat{\boldsymbol{\alpha}}\,])\mathbf{D}^T + \frac{1}{n}\,\mathbf{D}\,(n\,\mathrm{Cov}[\,\hat{\boldsymbol{\alpha}}\,])\,\mathbf{D}^T. \qquad \text{(B.8)}$$

V_1 and V_2 are defined as in Appendix A.4, and $\mathbf{C} = (C_1, ..., C_m)$ and $\mathbf{D} = (D_1, ..., D_m)$ are defined, for $k = 1, ..., m$, as

$$\begin{aligned}C_k = &\sum_{s=1}^{K} r_s\,\mathrm{Cov}[\,Y,\,X_k\,(1-p(\mathbf{X};\boldsymbol{\alpha}_o))\,|\,Z=1,\,S_{so}=1\,] \\ &+ \sum_{s=1}^{K} r_s\,\mathrm{Cov}[\,Y,\,X_k\,p(\mathbf{X};\boldsymbol{\alpha}_o)\,|\,Z=0,\,S_{so}=1\,], \qquad \text{(B.9)}\end{aligned}$$

$$D_k = \left.\frac{\partial\beta^*}{\partial\alpha_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} - \sum_{j=1}^{K-1}\left.\frac{\partial\beta^*}{\partial q_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} f_p(q_{jo})^{-1}\,\frac{\partial}{\partial\alpha_k}\,\{\mathbb{E}[\,1_{[p(\mathbf{X};\boldsymbol{\alpha})<q_j]}\,]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o},$$

and

$$\beta^* = \sum_{s=1}^{K} r_s \{\mathbb{E}[Y \mid Z = 1, S_s = 1] - \mathbb{E}[Y \mid Z = 0, S_s = 1]\}, \qquad \text{(B.10)}$$

which is equal to the 'true' value of $\hat{\beta}^s$, β_o^s, but seen as a function of the strata boundaries, $\mathbf{q}$, and the propensity score parameters, $\boldsymbol{\alpha}$, rather than evaluated at their population values. The covariance matrix $\text{Cov}[\hat{\boldsymbol{\alpha}}]$ is defined in terms of its inverse, for $j, k = 1, ..., m$, as

$$(n \, \text{Cov}[\hat{\boldsymbol{\alpha}}])^{-1}_{jk} = \mathbb{E}[p(\mathbf{X}; \boldsymbol{\alpha})(1 - p(\mathbf{X}; \boldsymbol{\alpha})) \, X_j \, X_k].$$

B.4.5 An alternative parameterization

We now re-parameterize the two last terms in (B.8) in order to simplify the estimation process. In order to do this we show that $\mathbf{D} = -\mathbf{C} + \mathbf{e}$, where $\mathbf{e} = (e_1, ..., e_m)$ is some error term. Then substituting this into (B.8) gives,

$$\mathbb{V}_e[\hat{\beta}^s] = V_1 + V_2 - \frac{1}{n} \mathbf{C} \, (n \, \text{Cov}[\hat{\boldsymbol{\alpha}}]) \, \mathbf{C}^T + \frac{1}{n} \mathbf{e} \, (n \, \text{Cov}[\hat{\boldsymbol{\alpha}}]) \, \mathbf{e}^T.$$

Then defining $V_3 = -\frac{1}{n} \mathbf{C} \, (n \, \text{Cov}[\hat{\boldsymbol{\alpha}}]) \, \mathbf{C}^T$ and $V_4 = \frac{1}{n} \mathbf{e} \, (n \, \text{Cov}[\hat{\boldsymbol{\alpha}}]) \, \mathbf{e}^T$,

$$\mathbb{V}_e[\hat{\beta}^s] = V_1 + V_2 + V_3 + V_4.$$

We begin by expressing the first term of $\mathbf{D}$, the derivative $\frac{\partial \beta^*}{\partial \boldsymbol{\alpha}}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$, as $-\mathbf{C}$ plus some error term which we will denote by $\mathbf{e}_{\boldsymbol{\alpha}} = (e_{\boldsymbol{\alpha}1}, ..., e_{\boldsymbol{\alpha}m})$. Using the formula

$$\mathbb{E}[Y \, Z \, S_s] = \mathbb{E}[Y \mid Z = 1, S_s = 1] \, \mathbb{E}[Z \, S_s],$$

we see that

$$\begin{aligned} &\frac{\partial}{\partial \alpha_k} \{\mathbb{E}[Y \mid Z = 1, S_s = 1]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \\ &= \frac{1}{d_{so}} \frac{\partial}{\partial \alpha_k} \{\mathbb{E}[Y \, Z \, S_s]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} - \frac{\mathbb{E}[Y \mid Z = 1, S_{so} = 1]}{d_{so}} \frac{\partial}{\partial \alpha_k} \{\mathbb{E}[Z \, S_s]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}. \qquad \text{(B.11)} \end{aligned}$$

A similar expression holds for the derivative of $\mathbb{E}[Y \mid Z = 0, S_s = 1]$. We now calculate the two derivatives in the right-hand side of (B.11), and give the analogous derivatives for the untreated group.

The derivatives $\frac{\partial}{\partial \boldsymbol{\alpha}}\{\mathbb{E}[\mathbf{Z}\,\mathbf{S_s}]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$ and $\frac{\partial}{\partial \boldsymbol{\alpha}}\{\mathbb{E}[(\mathbf{1}-\mathbf{Z})\,\mathbf{S_s}]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$

We show the calculation of only the first of these derivatives, since the second can be derived in exactly the same way. By conditioning on $\mathbf{X}$ and taking expectations over the 'true' distribution of the data, we can write, for $k = 1, ..., m$,

$$\frac{\partial}{\partial \alpha_k}\{\mathbb{E}[Z\,S_s]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \frac{\partial}{\partial \alpha_k}\{\mathbb{E}[p(\mathbf{X};\boldsymbol{\alpha}_o)\,S_s]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$$

$$= \frac{\partial}{\partial \alpha_k}\left\{\int_{q_{s-1}}^{q_s} \mathbb{E}[p(\mathbf{X};\boldsymbol{\alpha}_o)\,|\,p(\mathbf{X};\boldsymbol{\alpha}) = r]\; f_p(r;\boldsymbol{\alpha})\,dr\right\}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o},$$

where the function $f_p(\,\cdot\,;\boldsymbol{\alpha})$ is the probability density function of $p(\mathbf{X};\boldsymbol{\alpha})$, i.e. the probability density function of the propensity score seen as a function of $\boldsymbol{\alpha}$. Exchanging the order of integration and differentiation, and then differentiating using the product rule gives

$$\frac{\partial}{\partial \alpha_k}\{\mathbb{E}[ZS_s]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \int_{q_{s-1}}^{q_s} \frac{\partial}{\partial \alpha_k}\{\mathbb{E}[p(\mathbf{X};\boldsymbol{\alpha}_o)\,|\,p(\mathbf{X};\boldsymbol{\alpha}) = r]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\; f_p(r;\boldsymbol{\alpha}_o)\,dr$$

$$+ \int_{q_{s-1}}^{q_s} \mathbb{E}[p(\mathbf{X};\boldsymbol{\alpha}_o)\,|\,p(\mathbf{X};\boldsymbol{\alpha}_o) = r]\frac{\partial}{\partial \alpha_k}\{f_p(r;\boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\,dr.$$

In order to simplify this, we first consider the derivative of the conditional expectation, $\mathbb{E}[p(\mathbf{X};\boldsymbol{\alpha}_o)\,|\,p(\mathbf{X};\boldsymbol{\alpha}) = r]$. Differentiating model (B.1) shows that the derivative of $p(\mathbf{X};\boldsymbol{\alpha})$ with respect to α_k is $X_k\,p(\mathbf{X};\boldsymbol{\alpha})(1-p(\mathbf{X};\boldsymbol{\alpha}))$, for $k = 1, ..., m$. Then a Taylor series expansion of $p(\mathbf{X};\boldsymbol{\alpha}_o)$ gives,

$$p(\mathbf{X};\boldsymbol{\alpha}_o) = p(\mathbf{X};\boldsymbol{\alpha}) + \sum_{k=1}^{m}(\alpha_{ko}-\alpha_k)\;X_k\;p(\mathbf{X};\boldsymbol{\alpha})\;(1-p(\mathbf{X};\boldsymbol{\alpha})) + \sum_{k=1}^{m} O((\alpha_{ko}-\alpha_k)^2).$$

Taking expectations of this gives

$$\mathbb{E}[p(\mathbf{X};\boldsymbol{\alpha}_o)\,|\,p(\mathbf{X};\boldsymbol{\alpha}) = r] = r + \sum_{k=1}^{m}(\alpha_{ko}-\alpha_k)\;\mathbb{E}[X_k|\,p(\mathbf{X};\boldsymbol{\alpha}) = r]\,r\;(1-r)$$

$$+ \sum_{k=1}^{m} O_p((\alpha_{ko}-\alpha_k)^2).$$

Differentiating this with respect to α_k, for $k = 1, ..., m$, gives

$$\frac{\partial}{\partial \alpha_k}\{\mathbb{E}[p(\mathbf{X};\boldsymbol{\alpha}_o)\,|\,p(\mathbf{X};\boldsymbol{\alpha}) = r]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -r\,(1-r)\,\mathbb{E}[X_k|\,p(\mathbf{X};\boldsymbol{\alpha}) = r].$$

Note that this is an exact result since for $b \geq 2$,

$$\frac{\partial}{\partial \alpha_k}\{g(\boldsymbol{\alpha})(\alpha_{ko} - \alpha_k)^b\} = g'(\boldsymbol{\alpha}_o)(\alpha_{ko} - \alpha_{ko})^b + g(\boldsymbol{\alpha}_o)\, b\, (\alpha_{ko} - \alpha_{ko})^{(b-1)} = 0.$$

Then

$$\frac{\partial}{\partial \alpha_k}\, \{\mathbb{E}[\, Z S_s\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -\mathbb{E}\,[\, X_k\, p\,(\mathbf{X};\boldsymbol{\alpha}_o)(1 - p\,(\mathbf{X};\boldsymbol{\alpha}_o))\, S_{so}\,] + I_{f_1 k}$$

where $$I_{f_1 k} = \int_{q_{(s-1)o}}^{q_{so}} r \frac{\partial}{\partial \alpha_k}\, \{f_p(r;\boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\, dr.$$

Similarly,

$$\frac{\partial}{\partial \alpha_k}\, \{\mathbb{E}[\,(1 - Z) S_s\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \mathbb{E}\,[\, X_k\, p\,(\mathbf{X};\boldsymbol{\alpha}_o)(1 - p\,(\mathbf{x};\boldsymbol{\alpha}_o))\, S_{so}\,] + I_{f_0 k}$$

where $$I_{f_0 k} = \int_{q_{(s-1)o}}^{q_{so}} (1 - r) \frac{\partial}{\partial \alpha_k}\, \{f_p(r;\boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\, dr.$$

Note that the expectation that appears in both the derivatives above is negative in the first and positive in the second. This is due to the fact that

$$\frac{\partial\, p\,(\mathbf{X};\boldsymbol{\alpha})}{\partial \alpha_k} = -\frac{\partial\,(1 - p\,(\mathbf{X};\boldsymbol{\alpha}))}{\partial \alpha_k}.$$

The derivatives $\frac{\partial}{\partial \boldsymbol{\alpha}}\{\mathbb{E}\,[\,\mathbf{Y}\,\mathbf{Z}\,\mathbf{S_s}\,]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$ and $\frac{\partial}{\partial \boldsymbol{\alpha}}\{\mathbb{E}\,[\,\mathbf{Y}\,(\mathbf{1}-\mathbf{Z})\,\mathbf{S_s}\,]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$

As before, we calculate only the first of these derivatives, since the second can be derived in exactly the same way. By conditioning on $\mathbf{X}$ and taking expectations over the 'true' distribution of the data, we can write, for $k = 1, ..., m$,

$$\begin{aligned}
\frac{\partial}{\partial \alpha_k}\, \{\mathbb{E}\,[\, Y\, Z\, S_s\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} &= \frac{\partial}{\partial \alpha_k}\, \{\mathbb{E}\,[\, Y_1\, p\,(\mathbf{X};\boldsymbol{\alpha}_o)\, S_s\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \\
&= \int_{q_{(s-1)}}^{q_s} \frac{\partial}{\partial \alpha_k}\, \{\mathbb{E}\,[\, Y_1\, p\,(\mathbf{X};\boldsymbol{\alpha}_o)\,|\, p\,(\mathbf{X};\boldsymbol{\alpha}) = r\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\, f_p(r;\boldsymbol{\alpha}_o)\, dr \\
&+ \int_{q_{(s-1)}}^{q_s} \mathbb{E}\,[\, Y_1\, p\,(\mathbf{X};\boldsymbol{\alpha}_o)\,|\, p\,(\mathbf{X};\boldsymbol{\alpha}_o) = r\,]\, \frac{\partial}{\partial \alpha_k}\, \{\, f_p(r;\boldsymbol{\alpha})\,\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\, dr.
\end{aligned}$$

where, as before, $f_p(\,\cdot\,;\boldsymbol{\alpha})$ is the probability density function of the propensity score.

seen as a function of $\boldsymbol{\alpha}$, rather than evaluated at $\boldsymbol{\alpha}_o$. Letting $g(\mathbf{X}) = \mathbb{E}[Y_1 \,|\, \mathbf{X}]$, a Taylor series expansion gives

$$g(\mathbf{X})\, p(\mathbf{X};\boldsymbol{\alpha}_o) = g(\mathbf{X})\, p(\mathbf{X};\boldsymbol{\alpha}) + \sum_{k=1}^{m} (\alpha_{ko} - \alpha_k)\, g(\mathbf{X})\, X_k\, p(\mathbf{X};\boldsymbol{\alpha})\,(1 - p(\mathbf{X};\boldsymbol{\alpha})) + \sum_{k=1}^{m} O((\alpha_{ko} - \alpha_k)^2),$$

and so taking expectations gives

$$\mathbb{E}[Y_1\, p(\mathbf{X};\boldsymbol{\alpha}_o) \,|\, p(\mathbf{X};\boldsymbol{\alpha}) = r] \;=\; r\,\mathbb{E}[Y_1 \,|\, p(\mathbf{X};\boldsymbol{\alpha}) = r] + \sum_{k=1}^{m} (\alpha_{ko} - \alpha_k)\,\mathbb{E}[Y_1\, X_k \,|\, p(\mathbf{X};\boldsymbol{\alpha}) = r]\, r\,(1-r) + \sum_{k=1}^{m} O_p((\alpha_{ko} - \alpha_k)^2).$$

which, when differentiated with respect to α_k, gives

$$\frac{\partial}{\partial \alpha_k} \{\mathbb{E}[Y_1\, p(\mathbf{X};\boldsymbol{\alpha}_o) \,|\, p(\mathbf{X};\boldsymbol{\alpha}) = r]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = r\,\frac{\partial}{\partial \alpha_k} \{\mathbb{E}[Y_1 \,|\, p(\mathbf{X};\boldsymbol{\alpha}) = r]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} - \mathbb{E}[Y_1\, X_k \,|\, p(\mathbf{X};\boldsymbol{\alpha}) = r]\, r\,(1-r).$$

This is again an exact result since the error terms in the Taylor series expansion all vanish when the derivative is evaluated at $\boldsymbol{\alpha}_o$. Then,

$$\frac{\partial}{\partial \alpha_k} \{\mathbb{E}[Y\, Z\, S_s]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -\mathbb{E}[Y_1\, X_k\, p(\mathbf{X};\boldsymbol{\alpha}_o)\,(1 - p(\mathbf{X};\boldsymbol{\alpha}_o))\, S_{so}] + \int_{q_{s-1}}^{q_s} r\,\frac{\partial}{\partial \alpha_k} \{\mathbb{E}[Y_1 \,|\, p(\mathbf{X};\boldsymbol{\alpha}) = r]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\, f_p(r;\boldsymbol{\alpha}_o)\, dr + \int_{q_{s-1}}^{q_s} r\,\mathbb{E}[Y_1 \,|\, p(\mathbf{X};\boldsymbol{\alpha}_o) = r]\,\frac{\partial}{\partial \alpha_k} \{\, f_p(r;\boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\, dr.$$

Therefore, combining two of the terms above,

$$\frac{\partial}{\partial \alpha_k} \{\mathbb{E}[Y\, Z\, S_s]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -\,\mathbb{E}[Y_1\, X_k\, p(\mathbf{X};\boldsymbol{\alpha}_o)\,(1 - p(\mathbf{X};\boldsymbol{\alpha}_o))\, S_{so}] + I_{Y_1 k},$$

where $I_{Y_1 k} = \int_{q_{(s-1)}}^{q_s} r\,\frac{\partial}{\partial \alpha_k} \{\mathbb{E}[Y_1 \,|\, p(\mathbf{X};\boldsymbol{\alpha}) = r]\, f_p(r;\boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\, dr.$

Similarly,

$$\frac{\partial}{\partial \alpha_k}\{\mathbb{E}\left[Y\left(1-Z\right)S_s\right]\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -\,\mathbb{E}\left[Y_0\,X_k\,p\left(\mathbf{X};\boldsymbol{\alpha}_o\right)\left(1-p\left(\mathbf{X};\boldsymbol{\alpha}_o\right)\right)S_{so}\right] + I_{Y_0k}.$$

where $I_{Y_0k} = \int_{q_{(s-1)}}^{q_s} (1-r)\frac{\partial}{\partial \alpha_k}\left\{\mathbb{E}\left[Y_0 \,|\, p\left(\mathbf{X};\boldsymbol{\alpha}\right)=r\right] f_p(r;\boldsymbol{\alpha})\right\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\, dr.$

The derivative $\frac{\partial \beta^*}{\partial \boldsymbol{\alpha}}$ Substituting the two derivatives we have just calculated into (B.11), we find that

$$\frac{\partial}{\partial \alpha_k}\left\{\mathbb{E}\left[Y \,|\, Z=1,\, S_s=1\right]\right\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \frac{\left(I_{Y_1k} - \mathbb{E}\left[Y \,|\, Z=1, S_{so}=1\right] I_{f_1k}\right)}{d_{so}} - E_{1k},$$

and

$$\frac{\partial}{\partial \alpha_k}\left\{\mathbb{E}\left[Y \,|\, Z=0,\, S_s=1\right]\right\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \frac{\left(I_{Y_0k} - \mathbb{E}\left[Y \,|\, Z=0, S_{so}=1\right] I_{f_0k}\right)}{r_s - d_{so}} + E_{0k},$$

where

$$E_{1k} = \frac{\mathbb{E}\left[Y_1\,X_k\,p\left(\mathbf{X};\boldsymbol{\alpha}_o\right)\left(1-p\left(\mathbf{X};\boldsymbol{\alpha}_o\right)\right)S_{so}\right]}{d_{so}} - \frac{\mathbb{E}\left[Y \,|\, Z=1, S_{so}=1\right]\mathbb{E}\left[X_k\,p\left(\mathbf{X};\boldsymbol{\alpha}_o\right)\left(1-p\left(\mathbf{X};\boldsymbol{\alpha}_o\right)\right)S_{so}\right]}{d_{so}}.$$

and

$$E_{0k} = \frac{\mathbb{E}\left[Y_0\,X_k\,p\left(\mathbf{X};\boldsymbol{\alpha}_o\right)\left(1-p\left(\mathbf{X};\boldsymbol{\alpha}_o\right)\right)S_{so}\right]}{r_s - d_{so}} - \frac{\mathbb{E}\left[Y \,|\, Z=0, S_{so}=1\right]\mathbb{E}\left[X_k\,p\left(\mathbf{X};\boldsymbol{\alpha}_o\right)\left(1-p\left(\mathbf{X};\boldsymbol{\alpha}_o\right)\right)S_{so}\right]}{r_s - d_{so}}.$$

Then substituting these derivatives into β^* (B.10) we have

$$\begin{aligned}\frac{\partial \beta^*}{\partial \alpha_k}\Big|_{\theta=\theta_o} &= \sum_{s=1}^{K} r_s \frac{\left(I_{Y_1k} - \mathbb{E}\left[Y \,|\, Z=1, S_{so}=1\right] I_{f_1k}\right)}{d_{so}} \\ &- \sum_{s=1}^{K} r_s \frac{\left(I_{Y_0k} - \mathbb{E}\left[Y \,|\, Z=0, S_{so}=1\right] I_{f_0k}\right)}{r_s - d_{so}} \\ &- \sum_{s=1}^{K} r_s\{E_{1k} + E_{0k}\}. \end{aligned} \tag{B.12}$$

We now show that $\sum_{s=1}^{K} r_s\{E_{1k} + E_{0k}\} = C_k$, for $k = 1, ..., m$. We can write

$$\frac{\mathbb{E}\left[\, Y_1 \, X_k \, p\left(\mathbf{X}; \boldsymbol{\alpha}_o\right) \left(1 - p\left(\mathbf{X}; \boldsymbol{\alpha}_o\right)\right) S_{so} \,\right]}{d_{so}} = \mathbb{E}\left[\, Y \, X_k \left(1 - p\left(\mathbf{X}; \boldsymbol{\alpha}_o\right)\right) \,|\, Z = 1, \, S_{so} = 1 \,\right],$$

and

$$\frac{\mathbb{E}\left[\, X_k \, p\left(\mathbf{X}; \boldsymbol{\alpha}_o\right) \left(1 - p\left(\mathbf{X}; \boldsymbol{\alpha}_o\right)\right) S_{so} \,\right]}{d_{so}} = \mathbb{E}\left[\, X_k \left(1 - p\left(\mathbf{X}; \boldsymbol{\alpha}_o\right)\right) \,|\, Z = 1, \, S_{so} = 1 \,\right].$$

Using the formula $\mathrm{Cov}\left[\, A \, B \,\right] = \mathbb{E}\left[\, A \, B \,\right] - \mathbb{E}\left[\, A \,\right] \mathbb{E}\left[\, B \,\right]$, we find that

$$E_{1k} = \mathrm{Cov}\left[\, Y, \, X_k \left(1 - p\left(\mathbf{X}; \boldsymbol{\alpha}_o\right)\right) \,|\, Z = 1, \, S_{so} = 1 \,\right].$$

Similarly,

$$E_{0k} = \mathrm{Cov}\left[\, Y, \, X_k \, p\left(\mathbf{X}; \boldsymbol{\alpha}_o\right) \,|\, Z = 0, \, S_{so} = 1 \,\right].$$

Comparing these equations with the definition of $\mathbf{C}$ (B.9), we see that, as required, $\sum_{s=1}^{K} r_s\{E_{1k} + E_{0k}\} = C_k$, for $k = 1, ..., m$. Therefore, if we define

$$e_{\boldsymbol{\alpha}k} = \sum_{s=1}^{K} r_s \left\{ \frac{\left(I_{Y_1 k} - \mathbb{E}\left[\, Y \,|\, Z = 1, S_{so} = 1 \,\right] I_{f_1 k}\right)}{d_{so}} - \frac{\left(I_{Y_0 k} - \mathbb{E}\left[\, Y \,|\, Z = 0, S_{so} = 1 \,\right] I_{f_0 k}\right)}{r_s - d_{so}} \right\}$$

then from (B.12) we have that

$$\left.\frac{\partial \beta^*}{\partial \alpha_k}\right|_{\theta = \theta_o} = -C_k + e_{\boldsymbol{\alpha}k},$$

If we further define

$$e_{\mathbf{q}k} = -\sum_{j=1}^{K-1} \left.\frac{\partial \beta^*}{\partial q_j}\right|_{\theta = \theta_o} f_p(q_{jo})^{-1} \, \frac{\partial}{\partial \alpha_k} \left\{\mathbb{E}\left[\, S_j \,\right]\right\}|_{\boldsymbol{\theta} = \boldsymbol{\theta}_o},$$

then $D_k = -C_k + e_{\boldsymbol{\alpha}k} + e_{\mathbf{q}k}$. We then combine these two sources of error, and write $\mathbf{e}_k = e_{\boldsymbol{\alpha}k} + e_{\mathbf{q}k}$ where $\mathbf{e} = (e_1, .., e_m)$, to show that $\mathbf{D} = -\mathbf{C} + \mathbf{e}$, as desired.

B.4.6 The re-parameterized variance of the stratified treatment effect estimator

Using this new parameterization, the asymptotic variance of the stratified treatment effect estimator can be written as

$$\mathbb{V}_e[\, \hat{\beta}^s \,] = V_1 + V_2 + V_3 + V_4,$$

where V_1 and V_2 are defined as in Appendix A.4, and

$$V_3 = -\frac{1}{n}\,\mathbf{C}\,(n\,\mathrm{Cov}[\,\hat{\boldsymbol{\alpha}}\,])\,\mathbf{C}^T$$
$$V_4 = \quad\frac{1}{n}\,\mathbf{e}\,(n\,\mathrm{Cov}[\,\hat{\boldsymbol{\alpha}}\,])\,\mathbf{e}^T$$

where $\mathbf{C} = (C_1, ..., C_m)$ is defined, for $k = 1, ..., m$, as

$$C_k = \sum_{s=1}^{K} r_s\ \mathrm{Cov}[\,Y,\ X_k\,(1 - p_o(\mathbf{X}))\,|\,Z = 1,\ S_{so} = 1\,]$$
$$+ \quad \sum_{s=1}^{K} r_s\ \mathrm{Cov}[\,Y,\ X_k\,p_o(\mathbf{X})\,|\,Z = 0,\ S_{so} = 1\,].$$

The covariance matrix $(n\,\mathrm{Cov}[\,\hat{\boldsymbol{\alpha}}\,])$ is defined in terms of its inverse, for $j, k = 1, ..., m$, as

$$(n\ \mathrm{Cov}[\,\hat{\boldsymbol{\alpha}}\,])^{-1}_{jk} = \mathbb{E}\,[\,p_o(\mathbf{X})\,(1 - p_o(\mathbf{X}))\ X_j\ X_k\,].$$

Finally, $\mathbf{e} = (e_1, ..., e_m)$ where e_k is defined as $e_k = e_{\boldsymbol{\alpha}k} + e_{\mathbf{q}k}$, for $k = 1, ..., m$, with

$$e_{\boldsymbol{\alpha}k} = \sum_{s=1}^{k} r_s \left\{ \frac{(I_{Y_1 k} - \mathbb{E}\,[\,Y\,|\,Z = 1, S_{so} = 1\,]\,I_{f_1 k})}{d_{so}} - \frac{(I_{Y_0 k} - \mathbb{E}\,[\,Y\,|\,Z = 0, S_{so} = 1\,]\,I_{f_0 k})}{r_s - d_{so}} \right\},$$

where

$$I_{f_1 k} = \int_{q_{(s-1)o}}^{q_{so}} r \frac{\partial}{\partial \alpha_k}\,\{f_p(r;\boldsymbol{\alpha})\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\ dr$$

$$I_{f_0 k} = \int_{q_{(s-1)o}}^{q_{so}} (1 - r) \frac{\partial}{\partial \alpha_k}\,\{f_p(r;\boldsymbol{\alpha})\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\ dr$$

$$I_{Y_1 k} = \int_{q_{(s-1)o}}^{q_{so}} r \frac{\partial}{\partial \alpha_k}\,\{\mathbb{E}\,[\,Y_1\,|\,p\,(\mathbf{X};\boldsymbol{\alpha}) = r\,]\,f_p(r;\boldsymbol{\alpha})\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\ dr$$

$$I_{Y_0 k} = \int_{q_{(s-1)o}}^{q_{so}} (1 - r) \frac{\partial}{\partial \alpha_k}\,\{\mathbb{E}\,[\,Y_0\,|\,p\,(\mathbf{X};\boldsymbol{\alpha}) = r\,]\,f_p(r;\boldsymbol{\alpha})\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\ dr.$$

and

$$e_{\mathbf{q}k} = -\sum_{j=1}^{K-1} \frac{\partial \beta^*}{\partial q_j}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} f_p(q_{jo})^{-1}\ \frac{\partial}{\partial \alpha_k}\,\{\mathbb{E}\,[\,1_{[p\,(\mathbf{X};\boldsymbol{\alpha}) < q_j]}\,]\}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

We have now written the variance of the stratified treatment effect estimator, $\hat{\beta}^s$ as the sum of four variance components, where the first is the variance formula typically used in practice, and the remaining three terms are quadratic forms around positive definite matrices. The implications of this variance formula, and the four components of variance, are discussed in the main text (Section 3.4).

Appendix C

Application of the variance fomulæ to a hypothetical situation

The asymptotic variance of the stratified treatment effect estimator, $\hat{\beta}^s$, has been calculated assuming that the propensity score is: (i) a known function of the observed covariates; and (ii) estimated using a correctly specified logistic regression model. The resulting two variance formulæ, $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$, where the 'k' and 'e' represent the propensity score being known and estimated respectively, are given in Theorems 3.1 and 3.2. In Chapter 5 these variance formulæ are applied to several simple hypothetical situations. This appendix describes the mathematical calculation of these two variances for a simple hypothetical situation [1]. The calculations described in this chapter are implemented in the mathematical software *Mathematica* [112]. The programs that implement these calculations can be found in Appendix D.

We begin, in Section C.1, by giving the general form of the hypothetical situation considered. In Section C.2 we calculate the quantities contained in $\mathbb{V}_k[\hat{\beta}^s]$ — the variance of the stratified estimator of treatment effect when the propensity score is a known function of the covariates. We then calculate all the quantities contained in $\mathbb{V}_e[\hat{\beta}^s]$ — the variance of the stratified estimator of treatment effect when the propensity score is estimated — that have not previously been calculated in Section C.3.

This appendix uses the notation given in Section 3.1.1. In order to ease our way into the calculations, we briefly review the general hypothetical situation.

[1] In this appendix we calculate the 'true' values of $\mathbb{V}_k[\hat{\beta}^s]$ and $\mathbb{V}_e[\hat{\beta}^s]$. In Chapter 6 we describe how to estimate the two variances from a dataset, as would be necessary in practice.

C.1 The hypothetical situation

In the hypothetical examples considered in Chapter 5 there are two covariates — a binary covariate, X_1, and a continuous covariate, X_2, with a distribution that depends on X_1. The propensity score depends on both covariates. The outcome, Y, is continuous and depends on the covariate X_2 and treatment status, Z, only. The individual-level treatment effect is the same for each subject and is therefore equal to β_o, the population average causal treatment effect, β_o. Details are as follows:

$$\begin{aligned}
\text{Outcome:} \quad & Y = \gamma_0 + \gamma_2 X_2 + \beta_o Z + \epsilon, \quad \epsilon \sim N(0, \sigma_e^2). \\
\text{Propensity score:} \quad & \ln\left\{\frac{p_o(\mathbf{X})}{1 - p(\mathbf{X})}\right\} = \alpha_{0o} + \alpha_{1o} X_1 + \alpha_{2o} X_2. \\
\text{Covariates:} \quad & \mathbb{P}(X_1 = 0) = \pi_{x0}, \qquad \mathbb{P}(X_1 = 1) = 1 - \pi_{x0} = \pi_{x1}, \\
& X_2 \,|\, X_1 = 0 \sim N(\mu_0, \sigma_0^2), \quad X_2 \,|\, X_1 = 1 \sim N(\mu_1, \sigma_1^2).
\end{aligned}$$

Each of the hypothetical examples in Chapter 5 is a variation of this general situation. In this appendix we assume that K strata of equal size are used. The accompanying computer programs, in Appendix D, use the parameter values of hypothetical example (b), Section 5.2.2, where two strata of equal size are used. This greatly simplifies the presentation of the *Mathematica* code and can be easily generalised to K strata.

C.2 Calculating the variance when the propensity score is known

In order to calculate $\mathbb{V}_k\,[\,\hat{\beta}^s\,]$, the asymptotic variance of the stratified treatment effect estimator when the propensity score is a known function of the observed covariates, given in Theorem 3.1, we need to calculate the following population quantities:

(a) $f_p(\,\cdot\,)$ — the probability density function of the propensity score;

(b) $\mathbf{q}_o = (q_{1o}, \ldots, q_{(K-1)o})$ — the population strata boundaries;

(c) $\mathbf{d}_o = (d_{1o}, \ldots, d_{Ko})$ — the population probabilities of being treated and in each stratum;

(d) $\mathbb{V}\,[\,Y \,|\, Z = t, S_{so} = 1\,]$, for $t = 0, 1$ and $s = 1, \ldots, K$, — the population variance of the outcome, given treatment status and strata;

(e) $\frac{\partial}{\partial q_k}\{\mathbb{E}[Y \mid Z = t, S_s = 1]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$, for $t = 0, 1$, $s = 1, ..., K$, and $k = 1, ..., K-1$, — the derivative of the expected outcome given treatment status and strata, with respect to the strata boundaries.

These population quantities are calculated below in the order given.

C.2.1 The probability density function of the propensity score

Each of the quantities needed to calculate the required variances involves the population probability density function of the propensity score. This distribution is therefore now calculated by the Jacobian change of variables method using the known distributions of the two covariates, X_1 and X_2. We consider the one-one map,

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \rightarrow \begin{pmatrix} p \\ X_1 \end{pmatrix},$$

where p is the propensity score, defined by

$$p(\mathbf{X}) = \frac{\exp\{\alpha_{0o} + \alpha_{1o} X_1 + \alpha_{2o} X_2\}}{1 + \exp\{\alpha_{0o} + \alpha_{1o} X_1 + \alpha_{2o} X_2\}}. \tag{C.1}$$

Since $\frac{\partial X_1}{\partial X_1} = 1$ and $\frac{\partial X_1}{\partial X_2} = 0$, the Jacobian matrix of this transformation is

$$J = \begin{vmatrix} \frac{\partial p}{\partial X_1} & \frac{\partial p}{\partial X_2} \\ \frac{\partial X_1}{\partial X_1} & \frac{\partial X_1}{\partial X_2} \end{vmatrix} = \left| \frac{\partial p}{\partial X_2} \right| = |\alpha_{2o}|\, p\,(1-p).$$

Then the usual change of variables formula gives

$$f_{p,x_1}(p, X_1) = \frac{f_{x_1,x_2}(X_1, X_2)}{|\alpha_{2o}|\, p\,(1-p)}, \tag{C.2}$$

where $f_{x_1,x_2}(\cdot)$ denotes the joint probability density function of the covariates X_1 and X_2 and $f_{p,x_1}(\cdot)$ denotes the joint density function of X_1 and the propensity score. We want to calculate the density function of the propensity score alone, $f_p(\cdot)$. Since X_1 is a binary covariate,

$$f_p(p) = f_{p,x_1}(p, 0) + f_{p,x_1}(p, 1). \tag{C.3}$$

If p and X_1 are known, the value of X_2 can be obtained using (C.1). Given p, we denote the value taken by X_2 when $X_1 = 0$ by $w_0(p)$ and we denote the value taken by X_2 when $X_1 = 1$ by $w_1(p)$. Then substituting the densities from (C.2) into the

probability density function (C.3) we get

$$f_p(p) = \frac{f_{x_1,x_2}(0, w_0(p)) + f_{x_1,x_2}(1, w_1(p))}{|\alpha_{2o}|\ p\ (1-p)}.$$

Writing $f_{x_2|x_1=t}(\,\cdot\,)$ for the density function of X_2 given $X_1 = t$, for $t = 0, 1$, this becomes

$$f_p(p) = \frac{\mathbb{P}(X_1 = 0)\ f_{x_2|x_1=0}(w_0(p)) + \mathbb{P}(X_1 = 1)\ f_{x_2|x_1=1}(w_1(p))}{|\alpha_{2o}|\ p\ (1-p)},$$

To simplify this density, we write $v_0(p) = f_{x_2|x_1=0}(w_0(p))$ and $v_1(p) = f_{x_2|x_1=1}(w_1(p))$. Remembering that $\mathbb{P}(X_1 = 0) = \pi_{x0}$ and $\mathbb{P}(X_1 = 1) = \pi_{x1}$, the probability density function of the propensity score is then given by

$$f_p(p) = \frac{\pi_{x0}\, v_0(p) + \pi_{x1}\, v_1(p)}{|\alpha_{2o}|\, p\,(1-p)}, \tag{C.4}$$

where

$$v_0(p) = \exp\left\{-\frac{(w_0(p) - \mu_0)^2}{2\sigma_0^2}\right\} \Big/ \sqrt{2\pi\sigma_0^2} \qquad w_0(p) = \frac{\ln\left(\frac{p}{1-p}\right) - \alpha_{0o}}{\alpha_{2o}},$$

$$v_1(p) = \exp\left\{-\frac{(w_1(p) - \mu_1)^2}{2\sigma_1^2}\right\} \Big/ \sqrt{2\pi\sigma_1^2} \qquad w_1(p) = \frac{\ln\left(\frac{p}{1-p}\right) - \alpha_{0o} - \alpha_{1o}}{\alpha_{2o}}.$$

C.2.2 The population strata boundaries

Having calculated the probability density function of the propensity score we are now in a position to calculate the population strata boundaries, $\mathbf{q}_o = (q_{1o}, \ldots, q_{(K-1)o})$. Since K strata of equal size are being used, the j^{th} population strata boundary, q_{jo}, solves the equation $\Psi(q_{jo}) = 0$ where

$$\Psi(q_j) = \int_0^{q_j} f_p(p)\ dp - j/K,$$

for $j = 1, \ldots, K-1$. The root of this equation, q_{jo}, is found by numerical approximation methods using the function **FindRoot** in the software *Mathematica*. The program which calculates the population strata boundaries for hypothetical example (b) (Section 5.2.2) can be found in Appendix D.2.

C.2.3 The probability of being treated and in each stratum

We now calculate the population probabilities of being treated and in each stratum, $\mathbf{d}_o = (d_{1o}, ..., d_{Ko})$. The s^{th} component of this, d_{so}, is defined as

$$d_{so} = \mathbb{P}(Z = 1,\, S_{so} = 1) = \mathbb{E}\,[\,Z\,S_{so}\,] = \mathbb{E}\,[\,p\,(\mathbf{X})\,S_{so}\,],$$

where the last equality is obtained by conditioning on the observed covariates, $\mathbf{X}$. Remembering that $S_{so} = 1_{[\,q_{(s-1)o} \le p_o(\mathbf{X}) < q_{so}\,]}$ is an indicator for the s^{th} population stratum, we can write this expectation in integral form as

$$d_{so} = \int_{q_{(s-1)o}}^{q_{so}} p\; f_p\,(p)\; dp.$$

Using the population strata boundaries, $\mathbf{q}_o = (q_{1o}, ..., q_{(K-1)o})$, calculated previously these integrals are calculated using the numerical integration function **NIntegrate** in the software *Mathematica*.

C.2.4 The conditional expectation of the outcome given treatment status and strata

We now calculate the population conditional expectation of the outcome given treatment status and strata, $\mathbb{E}\,[\,Y \mid Z = t,\, S_{so} = 1\,]$, for $t = 0, 1$ and $s = 1, ..., K$. These conditional expectations will be used to calculate the conditional variances of the outcome given treatment status and strata, required to calculate the variances $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$. These conditional expectations are also needed to calculate the population stratified treatment effect, β_o^s. We saw in Chapter 3 that the population average causal treatment effect, β_o, is not usually equal to the population stratified treatment effect, β_o^s, where

$$\beta_o^s = \sum_{s=1}^{K} (\mathbb{E}\,[\,Y \mid Z = 1,\, S_{so} = 1\,] - \mathbb{E}\,[\,Y \mid Z = 0,\, S_{so} = 1\,])/K.$$

Therefore, as well as calculating the two variances for each hypothetical situation we also calculate the value of β_o^s in order to see how much residual confounding there is in each situation.

We calculate only the conditional expectation $\mathbb{E}[\,Y \mid Z = 1,\, S_{so} = 1\,]$. The conditional expectation $\mathbb{E}[\,Y \mid Z = 0,\, S_{so} = 1\,]$ can be calculated in the same way. By conditioning on the propensity score, we can write

$$\mathbb{E}[\,Y \mid Z = 1,\, S_{so} = 1\,] = \int_{q_{(s-1)o}}^{q_{so}} \mathbb{E}[\,Y \mid Z = 1, p_o(\mathbf{X}) = p\,]\; f(p \mid Z = 1,\, S_{so} = 1)\; dp. \tag{C.5}$$

The conditional density function can be written as

$$f(p \mid Z = 1,\, S_{so} = 1) = \frac{\mathbb{P}(Z = 1,\, S_{so} = 1 \mid p_o(\mathbf{X}) = p)\, f_p(p)}{\mathbb{P}(Z = 1,\, S_{so} = 1)} = \frac{p\, f_p(p)}{d_{so}}, \tag{C.6}$$

where $f_p(\,\cdot\,)$ is the probability density function of the propensity score and d_{so} is the population probability of being treated and in the s^{th} population stratum.

By conditioning on the covariates X_1 and X_2, the conditional expectation given the propensity score can be written as

$$\mathbb{E}[\,Y \mid Z = 1, p_o(\mathbf{X}) = p\,] = \int_{X_1, X_2} \mathbb{E}[\,Y \mid Z = 1, X_1,\, X_2\,]\; f(X_1, X_2 \mid p_o(\mathbf{X}) = p)\; dX_1\, dX_2.$$

In the notation introduced in Section C.2.1, the conditional density of X_1 and X_2 given the propensity score is

$$f(X_1,\, X_2 \mid p_o(X) = p) = \begin{cases} \frac{\pi_{x0}\, v_0(p)}{\pi_{x0}\, v_0(p) + \pi_{x1}\, v_1(p)} & \text{if } X_1 = 0 \text{ and } X_2 = w_0(p) \\ \frac{\pi_{x1}\, v_1(p)}{\pi_{x0}\, v_0(p) + \pi_{x1}\, v_1(p)} & \text{if } X_1 = 1 \text{ and } X_2 = w_1(p) \\ 0 & \text{otherwise.} \end{cases} \tag{C.7}$$

Referring to the known distribution of the outcome, Y, given in Section C.1,

$$\mathbb{E}[\,Y \mid Z = 1,\, X_1 = 0,\, X_2 = w_0(p)\,] = \gamma_0 + \beta_o + \gamma_2\, w_0(p)$$
$$\mathbb{E}[\,Y \mid Z = 1,\, X_1 = 1,\, X_2 = w_1(p)\,] = \gamma_0 + \beta_o + \gamma_2\, w_1(p),$$

and so

$$\mathbb{E}[\,Y \mid Z = 1,\, p\,] = \gamma_0 + \beta_o + \gamma_2 \left\{ \frac{\pi_{x0}\, v_0(p)\, w_0(p) + \pi_{x1}\, v_1(p)\, w_1(p)}{\pi_{x0}\, v_0(p) + \pi_{x1}\, v_1(p)} \right\}.$$

Then combining (C.5) and (C.6) and substituting in the expression for the probability density function of the propensity score (C.4), we obtain

$$\mathbb{E}[Y \mid Z = 1,\, S_{so} = 1] = \gamma_0 + \beta_o + \frac{\gamma_2}{d_{so}} \int_{q_{(s-1)o}}^{q_{so}} \frac{\{\pi_{x0}\, v_0(p)\, w_0(p) + \pi_{x1}\, v_1(p)\, w_1(p)\}}{|\alpha_{2o}|\,(1-p)}\, dp.$$

Similarly,

$$\mathbb{E}[Y \mid Z = 0,\, S_{so} = 1] = \gamma_0 + \frac{\gamma_2}{1/K - d_{so}} \int_{q_{(s-1)o}}^{q_{so}} \frac{\{\pi_{x0}\, v_0(p)\, w_0(p) + \pi_{x1}\, v_1(p)\, w_1(p)\}}{|\alpha_{2o}|\, p}\, dp.$$

C.2.5 The conditional variance of the outcome given treatment status and strata

The conditional variance given treatment status and strata is calculated using the equality

$$\mathbb{V}[Y \mid Z = t,\, S_{so} = 1] = \mathbb{E}[Y^2 \mid Z = t,\, S_{so} = 1] - (\mathbb{E}[Y \mid Z = t,\, S_{so} = 1])^2, \quad \text{(C.8)}$$

for $t = 0, 1$ and $s = 1, ..., K$. We have already calculated $\mathbb{E}[Y \mid Z = t,\, S_{so} = 1]$. The conditional expectations of Y^2 given treatment status and strata can be calculated in the same manner. Substituting these expectations into (C.8) gives the required variance. Recollect that the error term in the outcome, ϵ, has variance σ_e^2. Then for $s = 1, ..., K$,

$$\mathbb{V}[Y \mid Z = 1,\, S_{so} = 1] = \sigma_e^2 + A_{1s} - B_{1s}^2,$$
$$\mathbb{V}[Y \mid Z = 0,\, S_{so} = 1] = \sigma_e^2 + A_{0s} - B_{0s}^2,$$

with

$$A_{1s} = \frac{\gamma_2^2}{d_{so}} \int_{q_{(s-1)o}}^{q_{so}} \frac{\{\pi_{x0}\, v_0(p)\, w_0^2(p) + \pi_{x1}\, v_1(p)\, w_1^2(p)\}}{|\alpha_{2o}|\,(1-p)}\, dp,$$

$$B_{1s} = \frac{\gamma_2}{d_{so}} \int_{q_{(s-1)o}}^{q_{so}} \frac{\{\pi_{x0}\, v_0(p)\, w_0(p) + \pi_{x1}\, v_1(p)\, w_1(p)\}}{|\alpha_{2o}|\,(1-p)}\, dp,$$

and

$$A_{0s} = \frac{\gamma_2^2}{1/K - d_{so}} \int_{q_{(s-1)o}}^{q_{so}} \frac{\{\pi_{x0}\, v_0(p)\, w_0^2(p) + \pi_{x1}\, v_1(p)\, w_1^2(p)\}}{|\alpha_{2o}|\, p}\, dp,$$

$$B_{0s} = \frac{\gamma_2}{1/K - d_{so}} \int_{q_{(s-1)o}}^{q_{so}} \frac{\{\pi_{x0}\, v_0(p)\, w_0(p) + \pi_{x1}\, v_1(p)\, w_1(p)\}}{|\alpha_{2o}|\, p}\, dp.$$

C.2.6 The derivative of the conditional expectation of the outcome with respect to the strata boundaries

We now calculate the derivative of the conditional expectation of the outcome given treatment and strata, with respect to the strata boundaries,

$$\frac{\partial}{\partial q_k} \{\mathbb{E}[Y \mid Z = 1,\, S_s = 1]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o},$$

for $k = 1, ..., K-1$ and $s = 1, .., K$, where $S_s = 1_{[q_{(s-1)} \le p_o(\mathbf{X}) < q_s]}$ is an indicator for the s^{th} stratum defined by strata boundaries $q_{(s-1)}$ and q_s. Following the method used in Section C.2.4, but evaluating the integrals at strata boundaries $\mathbf{q} = (q_1, ..., q_{K-1})$, rather than at the population strata boundaries, $\mathbf{q}_o = (q_{1o}, ..., q_{(K-1)o})$, we see that

$$\mathbb{E}[Y \mid Z = 1,\, S_s = 1] = \gamma_0 + \beta_o + \frac{\gamma_2\, G_{1s}}{F_{1s}}. \tag{C.9}$$

where

$$F_{1s} = \int_{q_{s-1}}^{q_s} \frac{\{\pi_{x0}\, v_0(p) + \pi_{x1}\, v_1(p)\}}{|\alpha_{2o}|\,(1-p)}\, dp,$$

and

$$G_{1s} = \int_{q_{s-1}}^{q_s} \frac{\{\pi_{x0}\, v_0(p)\, w_0(p) + \pi_{x1}\, v_1(p)\, w_1(p)\}}{|\alpha_{2o}|\,(1-p)}\, dp.$$

When these two integrals are evaluated at the population strata boundaries, $\mathbf{q}_o$, we call them F_{1so} and G_{1so}. Then F_{1so} is equal to d_{so}, the population probability of being treated and in the s^{th} population strata. Using the quotient rule to differentiate (C.9) with respect to the strata boundaries gives

$$\frac{\partial}{\partial q_k} \{\mathbb{E}[Y \mid Z = 1,\, S_s = 1]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \frac{\gamma_2}{d_{so}} \left.\frac{\partial (G_{1s})}{\partial q_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} - \frac{\gamma_2\, G_{1so}}{d_{so}^2} \left.\frac{\partial (F_{1s})}{\partial q_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}. \tag{C.10}$$

Differentiating the two integrals, F_{1s} and G_{1s} with respect to the strata boundary q_k, for $k = 1, ..., K-1$, using the fundamental theorem of calculus, we obtain

$$\left.\frac{\partial\,(F_{1s})}{\partial q_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \begin{cases} \frac{\pi_{x0}\,v_0(q_{ko})+\pi_{x1}\,v_1(q_{ko})}{|\alpha_{2o}|\,(1-q_{so})} & \text{if } k = s, \\ -\frac{0.6\,v_0(q_{ko})+0.4\,v_1(q_{ko})}{|\alpha_{2o}|\,(1-q_{ko})} & \text{if } k = s-1, \\ 0 & \text{otherwise,} \end{cases}$$

$$\left.\frac{\partial\,(G_{1s})}{\partial q_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \begin{cases} \frac{\pi_{x0}\,v_0(q_{ko})\,w_0(q_{ko})+\pi_{x1}\,v_1(q_{ko})\,w_1(q_{ko})}{|\alpha_{2o}|\,(1-q_{ko})} & \text{if } k = s, \\ -\frac{\pi_{x0}\,v_0(q_{ko})\,w_0(q_{ko})+\pi_{x1}\,v_1(q_{ko})\,w_1(q_{ko})}{|\alpha_{2o}|\,(1-q_{ko})} & \text{if } k = s-1, \\ 0 & \text{otherwise.} \end{cases}$$

These two derivatives can then be substituted into (C.10) to obtain the required derivative.

Similarly, we can write the expectation of the outcome, given no treatment and strata as

$$\mathbb{E}\left[\,Y \mid Z = 0,\, S_s = 1\,\right] = \gamma_0 + \frac{\gamma_2\,G_{0s}}{F_{0s}}.$$

where

$$F_{0s} = \int_{q_{s-1}}^{q_s} \frac{\{\pi_{x0}\,v_0(p) + \pi_{x1}\,v_1(p)\}}{|\alpha_{2o}|\,p}\,dp,$$

$$\text{and} \qquad G_{0s} = \int_{q_{s-1}}^{q_s} \frac{\{\pi_{x0}\,v_0(p)\,w_0(p) + \pi_{x1}\,v_1(p)\,w_1(p)\}}{|\alpha_{2o}|\,p}\,dp.$$

When these integrals are evaluated at the population strata boundaries, $\mathbf{q}_o$, we call them F_{0so} and G_{0so}. Then F_{0so} is equal to $1/K - d_{so}$, the population probability of being untreated and in the s^{th} population strata. Using the quotient rule to differentiate the conditional expectation with respect to the strata boundaries gives

$$\begin{aligned} \left.\frac{\partial}{\partial q_k}\left\{\mathbb{E}\left[\,Y \mid Z = 0,\, S_s = 1\,\right]\right\}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = {} & \frac{\gamma_2}{1/K - d_{so}}\left.\frac{\partial\,(G_{0s})}{\partial q_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \\ & - \frac{\gamma_2\,G_{0so}}{(1/K - d_{so})^2}\left.\frac{\partial\,(F_{0s})}{\partial q_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}. \end{aligned} \tag{C.11}$$

Differentiating these two integrals with respect to the strata boundaries gives

$$\left.\frac{\partial\,(F_{0s})}{\partial q_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \begin{cases} \frac{\pi_{x0}\,v_0(q_{ko})+\pi_{x1}\,v_1(q_{ko})}{|\alpha_{2o}|\,q_{so}} & \text{if } k=s, \\ -\frac{\pi_{x0}\,v_0(q_{ko})+\pi_{x1}\,v_1(q_{ko})}{|\alpha_{2o}|\,q_{ko}} & \text{if } k=s-1. \\ 0 & \text{otherwise,} \end{cases}$$

$$\left.\frac{\partial\,(G_{0s})}{\partial q_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \begin{cases} \frac{\pi_{x0}\,v_0(q_{ko})\,w_0(q_{ko})+\pi_{x1}\,v_1(q_{ko})\,w_1(q_{ko})}{|\alpha_{2o}|\,q_{ko}} & \text{if } k=s, \\ -\frac{\pi_{x0}\,v_0(q_{ko})\,w_0(q_{ko})+\pi_{x1}\,v_1(q_{ko})\,w_1(q_{ko})}{|\alpha_{2o}|\,q_{ko}} & \text{if } k=s-1, \\ 0 & \text{otherwise.} \end{cases}$$

These two derivatives can then be substituted into (C.11) to obtain the required derivative.

C.3 Calculating the variance when the propensity score is estimated

In order to calculate $\mathbb{V}_e\,[\,\hat{\beta}^s\,]$, the asymptotic variance of the stratified estimator of treatment effect when the propensity score is estimated using a correctly specified logistic regression model, given in Theorem 3.2, we also need to calculate the following quantities:

(a) $\mathbb{E}\,[\,p_o(\mathbf{X})\,(1-p_o(\mathbf{X}))\,\mathbf{X}^T\,\mathbf{X}\,]$ — the (inverse of the) asymptotic covariance matrix of the estimated propensity score parameters;

(b) $\mathrm{Cov}[\,Y,\,X_k\,(1-p_o(\mathbf{X}))\,|\,Z=1,\,S_{so}=1\,]$, $\mathrm{Cov}[\,Y,\,\mathbf{X}\,p_o(\mathbf{X})\,|\,Z=0,\,S_{so}=1\,]$, for $s=1,\ldots,K$ and $k=0,1,2,$ — the covariance of the outcome and the covariates weighted by a function of the propensity score;

(c) $\frac{\partial}{\partial\alpha_k}\,\{\mathbb{E}\,[\,1[\,p\,(\mathbf{X})<q_j\,]\,]\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}$, for $j=1,\ldots,K-1$, and $k=0,1,2,$ — the derivative of the cumulative distribution of the propensity score with respect to the propensity score parameters;

(d) The integrals $I_{f_1k}, I_{f_ok}, I_{Y_1k}$ and I_{Y_0k} — integrals in the error term involving derivatives with respect to the propensity score parameters.

These quantities are calculated below in the order given.

C.3.1 The covariance matrix for the propensity score parameters

In order to calculate the asymptotic covariance matrix of the propensity score parameters, $(n\,\mathrm{Cov}[\,\hat{\boldsymbol{\alpha}}\,])$, it is necessary to evaluate several integrals of the form

$$\begin{aligned}
&\mathbb{E}\left[\, g(X_1, X_2)\, p_o(\mathbf{X})\,(1 - p_o(\mathbf{X}))\,\right] \\
&= \int_0^1 \int_{X_1, X_2} g(X_1, X_2)\, p\,(1-p)\, f(X_1, X_2 | p_o(\mathbf{X}) = p)\, f_p(p)\; dp\; dX_1\; dX_2,
\end{aligned}$$

where $f_p(p)$ is the probability density function of the propensity score and $g(\cdot)$ is some function of the two covariates, X_1 and X_2. If we write

$$\begin{aligned}
g(X_1 = 0,\, X_2 = w_0(p)\,) &= g_0(p) \\
g(X_1 = 1,\, X_2 = w_1(p)\,) &= g_1(p),
\end{aligned}$$

then substituting in the conditional density of the covariates given the propensity score (C.7) and the probability density function of the propensity score (C.4) gives

$$\mathbb{E}\left[\, g(X_1, X_2)\, p_o(\mathbf{X})\,(1 - p_o(\mathbf{X}))\,\right] = \int_0^1 \frac{\{\pi_{x0}\, v_0(p)\, g_0(p) + \pi_{x1}\, v_1(p)\, g_1(p)\}}{|\alpha_{2o}|}\; dp.$$

Each entry of the inverse of the asymptotic covariance matrix of the estimated propensity score parameters can be calculated using this formula.

C.3.2 The covariances of outcome, covariates and the propensity score

We wish to calculate the quantities

$$\begin{aligned}
&\mathrm{Cov}\left[\, Y,\, X_k\,(1 - p_o(\mathbf{X}))\,|\, Z = 1,\, S_{so} = 1\,\right] \\
&\mathrm{Cov}\left[\, Y,\, X_k\, p_o(\mathbf{X})\,|\, Z = 0,\, S_{so} = 1\,\right],
\end{aligned}$$

for $s = 1, \ldots, K$, and $k = 0, 1, 2$, where the covariate X_0 is defined as an intercept vector of 1's. We now calculate the first of these covariances. The second is obtained in the same way.

Since $\mathrm{Cov}[A, B] = \mathbb{E}[AB] - \mathbb{E}[A]\mathbb{E}[B]$, and we have already calculated the expectation $\mathbb{E}[Y \mid Z = 1, S_{so} = 1]$, we only need to calculate the two expectations.

$$\mathbb{E}[X_k(1 - p_o(\mathbf{X})) \mid Z = 1, S_{so} = 1]$$
$$\mathbb{E}[YX_k(1 - p_o(\mathbf{X})) \mid Z = 1, S_{so} = 1].$$

Following the method used in Section C.2.4, we obtain

$$\mathbb{E}[X_k(1 - p_o(\mathbf{X})) \mid Z = 1, S_{so} = 1] = \frac{1}{d_{so}} \int_{q_{(s-1)o}}^{q_{so}} \frac{\pi_{x0}\, v_0(p)\, l_0(p) + \pi_{x1}\, v_1(p)\, l_1(p)}{|\alpha_{2o}|}\, dp,$$

$$\mathbb{E}[X_k p_o(\mathbf{X}) \mid Z = 0, S_{so} = 1] = \frac{1}{1/K - d_{so}} \int_{q_{(s-1)o}}^{q_{so}} \frac{\pi_{x0}\, v_0(p)\, l_0(p) + \pi_{x1}\, v_1(p)\, l_1(p)}{|\alpha_{2o}|}\, dp,$$

where

$$l_0(p) = \mathbb{E}[X_k \mid X_1 = 0, X_2 = w_0(p)],$$
$$l_1(p) = \mathbb{E}[X_k \mid X_1 = 1, X_2 = w_1(p)].$$

And in the same way,

$$\mathbb{E}[Y X_k(1 - p_o(\mathbf{X})) \mid Z = 1, S_{so} = 1]$$
$$= \frac{1}{d_{so}} \int_{q_{(s-1)o}}^{q_{so}} \frac{\pi_{x0}\, v_0(p)\, h(1,0,p)\, l_0(p) + \pi_{x1}\, v_1(p)\, h(1,1,p)\, l_1(p)}{|\alpha_{2o}|}\, dp,$$
$$\mathbb{E}[Y X_k p_o(\mathbf{X}) \mid Z = 0, S_{so} = 1]$$
$$= \frac{1}{d_{so}} \int_{q_{(s-1)o}}^{q_{so}} \frac{\pi_{x0}\, v_0(p)\, h(0,0,p)\, l_0(p) + \pi_{x1}\, v_1(p)\, h(0,1,p)\, l_1(p)}{|\alpha_{2o}|}\, dp.$$

where for $r, t = 0, 1$, $h(t, r, p) = \mathbb{E}[Y \mid Z = t, X_1 = r, X_2 = w_r(p)]$.

C.3.3 The derivative of the cumulative density function of the propensity score with respect to the propensity score parameters

We now calculate the derivative of the cumulative distribution function of the propensity score, with respect to the propensity score parameters,

$$\frac{\partial}{\partial \alpha_k} \left\{ \mathbb{E}\left[1_{[p(\mathbf{X};\boldsymbol{\alpha}) < q_j]} \right] \right\} \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}_o},$$

for $k = 0, 1, 2$, and $j = 1, ..., K-1$. Writing $F_p(p; \boldsymbol{\alpha})$ for the cumulative density function of the propensity score, seen as a function of $\boldsymbol{\alpha}$, we can write this as

$$\frac{\partial}{\partial \alpha_k} \{F_p(p; \boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

Previously in this appendix, all quantities have been evaluated at the population strata boundaries, $\boldsymbol{\alpha}_o$. In the remaining two sections, we wish to differentiate with respect to $\boldsymbol{\alpha}$ and so we need to be clear about which functions depend on $\boldsymbol{\alpha}$. Therefore, we will now state dependence on $\boldsymbol{\alpha}$ explicitly. For example, we will write $p(\mathbf{X}; \boldsymbol{\alpha})$ in order to remind ourselves that the propensity score is a function of the unknown parameter $\boldsymbol{\alpha}$.

Following the argument of Section C.2.1, we see that whatever value $\boldsymbol{\alpha}$ takes, the function

$$p(\mathbf{X}; \boldsymbol{\alpha}) = \frac{\exp\{\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2\}}{1 + \exp\{\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2\}},$$

has probability density function

$$f_{p;\alpha}(p) = \frac{\pi_{x0}\, v_0(p\,; \boldsymbol{\alpha}) + \pi_{x1}\, v_1(p\,; \boldsymbol{\alpha})}{|\alpha_2|\; p\;(1-p)},$$

where

$$v_0(p; \boldsymbol{\alpha}) = \exp\left\{-\frac{(w_0(p) - \mu_0)^2}{2\sigma_0^2}\right\} \Big/ \sqrt{2\pi\sigma_0^2} \qquad w_0(p; \boldsymbol{\alpha}) = \frac{\ln\left(\frac{p}{1-p}\right) - \alpha_0}{\alpha_2},$$

$$v_1(p; \boldsymbol{\alpha}) = \exp\left\{-\frac{(w_1(p) - \mu_1)^2}{2\sigma_1^2}\right\} \Big/ \sqrt{2\pi\sigma_1^2} \qquad w_1(p; \boldsymbol{\alpha}) = \frac{\ln\left(\frac{p}{1-p}\right) - \alpha_0 - \alpha_1}{\alpha_2}. \tag{C.12}$$

We wish to calculate the derivative

$$\frac{\partial}{\partial \alpha_k} \{F_p(p; \boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \frac{\partial}{\partial \alpha_k} \left(\int_0^{q_j} \frac{\{\pi_{x0}\, v_0(p\,; \boldsymbol{\alpha}) + \pi_{x1}\, v_1(p\,; \boldsymbol{\alpha})\}}{|\alpha_2|\, p\,(1-p)}\, dp \right) \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

The order of differentiation and integration can be interchanged to get

$$\frac{\partial}{\partial \alpha_k}\ \{F_p(p;\boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \int_0^{q_j} \frac{\left\{\pi_{x0}\ \frac{\partial v_0(p\,;\boldsymbol{\alpha})}{\partial \alpha_k}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} + \pi_{x1}\ \frac{\partial v_1(p\,;\boldsymbol{\alpha})}{\partial \alpha_k}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\right\}}{|\alpha_{2o}|\, p\,(1-p)}\ dp$$
$$- 1_{[k=2]} \int_0^{q_j} \frac{\{\pi_{x0}\, v_0(p\,;\boldsymbol{\alpha}_o) + \pi_{x1}\, v_1(p\,;\boldsymbol{\alpha}_o)\}}{\alpha_{2o}^2\, p\,(1-p)}\ dp. \qquad \text{(C.13)}$$

Differentiating the functions $v_1(\,\cdot\,)$ and $v_0(\,\cdot\,)$ defined above (C.12) with respect to $\boldsymbol{\alpha}$ and evaluating them at their population values, $\boldsymbol{\alpha}_o$, gives

$$\frac{\partial v_0(p\,;\boldsymbol{\alpha})}{\partial \alpha_0} = \frac{v_0(p\,;\alpha_o)\ (w_0(p\,;\alpha_o)-\mu_0)}{\alpha_{2o}\ \sigma_0^2}, \qquad \frac{\partial v_1(p\,;\boldsymbol{\alpha})}{\partial \alpha_0} = \frac{v_1(p\,;\alpha_o)\ (w_1(p\,;\alpha_o)-\mu_1)}{\alpha_{2o}\ \sigma_1^2},$$
$$\frac{\partial v_0(p\,;\boldsymbol{\alpha})}{\partial \alpha_1} = 0, \qquad \frac{\partial v_1(p\,;\boldsymbol{\alpha})}{\partial \alpha_1} = \frac{v_1(p\,;\alpha_o)\ (w_1(p\,;\alpha_o)-\mu_1)}{\alpha_{2o}\ \sigma_1^2},$$
$$\frac{\partial v_0(p\,;\boldsymbol{\alpha})}{\partial \alpha_2} = \frac{v_0(p\,;\alpha_o)\ w_0(p\,;\alpha_o)\ (w_0(p\,;\alpha_o)-\mu_0)}{\alpha_{2o}\ \sigma_0^2}, \qquad \frac{\partial v_1(p\,;\boldsymbol{\alpha})}{\partial \alpha_2} = \frac{v_1(p\,;\alpha_o)\ w_1(p\,;\alpha_o)\ (w_1(p\,;\alpha_o)-\mu_1)}{\alpha_{2o}\ \sigma_1^2}.$$

Substituting these derivatives into (C.13) gives the required derivatives.

C.3.4 The integrals I_{f_1k}, I_{f_0k}, I_{Y_1k} and I_{Y_0k}

We begin by calculating the integrals I_{f_1k} and I_{f_0k}. The first of these is defined as

$$I_{f_1k} = \int_{q_{(s-1)o}}^{q_{so}} p \frac{\partial}{\partial \alpha_k}\ \{f_p(p;\boldsymbol{\alpha})\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\ dp.$$

We have already calculated the probability density function of the propensity score, as a function of the unknown parameters, $\boldsymbol{\alpha}$, (C.12). We merely substitute this into the definition above to get

$$I_{f_1k} = \int_{q_{(s-1)o}}^{q_{so}} \frac{\pi_{x0}\ \frac{\partial v_0(p\,;\boldsymbol{\alpha})}{\partial \alpha_k}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} + \pi_{x1}\ \frac{\partial v_1(p\,;\boldsymbol{\alpha})}{\partial \alpha_k}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}}{\alpha_{2o}(1-p)}\ dp$$
$$- 1_{[k=2]} \int_{q_{(s-1)o}}^{q_{so}} \frac{\pi_{x0}\, v_0(p;\boldsymbol{\alpha}_o) + \pi_{x1}\, v_1(p;\boldsymbol{\alpha}_o)}{\alpha_{2o}^2(1-p)}\ dr.$$

Substituting the derivatives of $v_0(\,\cdot\,)$ and $v_1(\,\cdot\,)$ above gives the required derivatives.

Similarly,

$$I_{f0k} = \int_{q_{(s-1)o}}^{q_{so}} \frac{\pi_{x0} \left.\frac{\partial v_0(p;\boldsymbol{\alpha})}{\partial \alpha_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} + \pi_{x1} \left.\frac{\partial v_1(p;\boldsymbol{\alpha})}{\partial \alpha_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}}{\alpha_{2o} p}\, dp$$
$$- 1_{[k=2]} \int_{q_{(s-1)o}}^{q_{so}} \frac{\pi_{x0}\, v_0(p;\boldsymbol{\alpha}_o) + \pi_{x1}\, v_1(p;\boldsymbol{\alpha}_o)}{\alpha_{2o}^2 p}\, dr.$$

The conditional expectations $\mathbb{E}[\,\mathbf{Y_1} \,|\, \mathbf{p(X;\alpha) = p}\,]$ and $\mathbb{E}[\,\mathbf{Y_0} \,|\, \mathbf{p(X;\alpha) = p}\,]$

The first of these conditional expectation can be written as

$$\mathbb{E}[\,Y_1 \,|\, p(\mathbf{X};\boldsymbol{\alpha}) = p\,] = \int_{X_1,X_2} \mathbb{E}[\,Y_1 \,|\, X_1,\, X_2\,]\, f(X_1, X_2 \,|\, p(\mathbf{X};\boldsymbol{\alpha}) = p)\, dX_1\, dX_2.$$

The conditional density of the covariates given the propensity score — evaluated at $\boldsymbol{\alpha}$ rather than at the population values $\boldsymbol{\alpha}_o$ — is

$$f(X_1,\, X_2 \,|\, p(\mathbf{X};\boldsymbol{\alpha}) = p) = \begin{cases} \frac{\pi_{x0}\, v_0(p;\boldsymbol{\alpha})}{\pi_{x0}\, v_0(p;\boldsymbol{\alpha}) + \pi_{x1}\, v_1(p;\boldsymbol{\alpha})} & \text{if } X_1 = 0 \text{ and } X_2 = w_0(p;\boldsymbol{\alpha}) \\ \frac{\pi_{x1}\, v_1(p;\boldsymbol{\alpha})}{\pi_{x0}\, v_0(p;\boldsymbol{\alpha}) + \pi_{x1}\, v_1(p;\boldsymbol{\alpha})} & \text{if } X_1 = 1 \text{ and } X_2 = w_1(p;\boldsymbol{\alpha}) \\ 0 & \text{otherwise.} \end{cases}$$

Recollect that we defined $r, t = 0, 1,\ \ h(t, r, p) = \mathbb{E}[\,Y \,|\, Z = t,\, X_1 = r,\, X_2 = w_r(p)\,]$.
Then

$$\mathbb{E}[\,Y_1 \,|\, X_1 = 0,\, X_2 = w_0(p;\boldsymbol{\alpha})\,] = h(1, 0, p)$$
$$\mathbb{E}[\,Y_0 \,|\, X_1 = 1,\, X_2 = w_1(p;\boldsymbol{\alpha})\,] = h(1, 1, p).$$

Then

$$\mathbb{E}[\,Y_1 \,|\, p(\mathbf{X};\boldsymbol{\alpha}) = p\,] = \frac{\pi_{x0}\, v_0(p;\boldsymbol{\alpha}) h(1, 0, p) + \pi_{x1}\, v_1(p;\boldsymbol{\alpha}) h(1, 1, p)}{\pi_{x0}\, v_0(p;\boldsymbol{\alpha}) + \pi_{x1}\, v_1(p;\boldsymbol{\alpha})}.$$

Similarly,

$$\mathbb{E}[\,Y_0 \,|\, p(\mathbf{X};\boldsymbol{\alpha}) = p\,] = \frac{\pi_{x0}\, v_0(p;\boldsymbol{\alpha}) h(0, 0, p) + \pi_{x1}\, v_1(p;\boldsymbol{\alpha}) h(0, 1, p)}{\pi_{x0}\, v_0(p;\boldsymbol{\alpha}) + \pi_{x1}\, v_1(p;\boldsymbol{\alpha})}.$$

The integrals I_{Y_1k} and I_{Y_0k}

These integrals are defined as

$$I_{Y_1k} = \int_{q_{(s-1)o}}^{q_{so}} p\frac{\partial}{\partial\alpha_k}\left\{\mathbb{E}\left[Y_1 \,|\, p\left(\mathbf{X};\boldsymbol{\alpha}\right) = p\right] f_p(p;\boldsymbol{\alpha})\right\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\, dr$$

$$I_{Y_0k} = \int_{q_{(s-1)o}}^{q_{so}} (1-p)\frac{\partial}{\partial\alpha_k}\left\{\mathbb{E}\left[Y_0 \,|\, p\left(\mathbf{X};\boldsymbol{\alpha}\right) = p\right] f_p(p;\boldsymbol{\alpha})\right\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}\, dr.$$

We have calculated the conditional expectations and the probability density function in these two integrals. These can be differentiated with respect to α_k separately using the product rule. We then merely differentiate I_{Y_1k} and I_{Y_0k} using the product rule and integrate over the calculated derivatives in order to obtain the integrals I_{Y_1k} and I_{Y_0k}.

We have now described the calculation of all quantities involved in the two variance formulae $\mathbb{V}_k\left[\hat{\beta}^s\right]$ and $\mathbb{V}_e\left[\hat{\beta}^s\right]$. The accompanying programs that implement these calculations using the software *Mathematica* can be found in Appendix D.2 and Appendix D.3.

Appendix D

Appendix D: Computer programs

This appendix contains a selection of the computer programs used during the course of this thesis. The asymptotic variance of the stratified estimator of treatment effect was calculated assuming that the propensity score is: (i) a known function of the observed covariates; and (ii) estimated using a correctly specified logistic regression model. The resulting variances are referred to as $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$ respectively. Simulation studies were performed for a simple hypothetical situation in order to obtain empirical estimates of these two variances. The 'true' values of the variances were also calculated in order to compare the empirical and theoretical variances. This appendix contains the computer programs used to carry out these simulation studies. These were carried out using the software programs *Stata* [99] and *Mathematica* [112].

We begin by briefly reviewing the hypothetical situation that the programs use. Appendix D.1 contains the *Stata* program used to simulate datasets, Appendix D.2 contains the *Mathematica* program used to obtain the population strata boundaries, and Appendix D.3 contains the *Mathematica* program used to calculate the variances $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$.

A star at the beginning of a line is used to indicate a comment and three forward slashes /// are used to indicate a single line broken into two in order to simplify the presentation.

The hypothetical situation

The hypothetical situation considered in this appendix is example (b) described in Chapter 5. There are two covariates – a binary covariate, X_1, and a continuous covariate, X_2, with a distribution that depends on X_1. The propensity score depends on both covariates. The outcome, Y, is continuous and depends on the covariate X_2 and treatment status, Z, only. The causal effect of treatment is the same for each

subject and is therefore equal to β_o, the population causal effect of treatment. Two equal-sized strata were used for this example. This makes the computer programs more legible. Extending them to use K strata is straightforward. Details of the hypothetical situation are as follows:

$$\begin{aligned}
\text{Outcome:} \qquad & Y = 35 + 0.3\, X_2 + 2\, Z + \epsilon, \quad \epsilon \sim N(0, 10^2). \\
\text{Propensity score:} \qquad & \ln\left(\frac{p}{1-p}\right) = -3 - 0.5\, X_1 + 0.05\, X_2. \\
\text{Covariates:} \qquad & \mathbb{P}(X_1 = 0) = 0.6, \\
& X_2 \,|\, X_1 = 0 \sim N(70, 10^2); \quad X_2 \,|\, X_1 = 1 \sim N(60, 15^2).
\end{aligned}$$

A simulation study was performed for this hypothetical situation, using the software *Stata*. 3,000 datasets were simulated, each containing $n = 2{,}000$ subjects. Each simulated dataset was used to estimate the stratified estimator of treatment effect, $\hat{\beta}^s$, in two ways: (i) stratifying by the true propensity score; and (ii) stratifying by the estimated propensity score which was estimated using a correctly specified logistic regression model. An empirical estimate of the variance of $\hat{\beta}^s$ is then obtained, assuming that the propensity score is (i) known and (ii) estimated by taking the sample variance of the 3,000 estimates of treatment effect. The 'true' asymptotic variances, $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and $\mathbb{V}_e[\,\hat{\beta}^s\,]$, were then calculated using *Mathematica*. Details of the calculation are given in Appendix C. This simulation study was repeated for several hypothetical situations, each of which was a variation on the above situation.

D.1 Stata program used to obtain empirical estimates of the variances

```
*****************************************
* 0. Parameters that define the problem  *
*****************************************

* Individual-level treatment effect
global b0 = 2

* Outcome parameters
global g0  = 35
global g2  = 0.3
global sde = 10

* Propensity score parameters
global a0 = -3
global a1 = -0.5
global a2 =  0.05

* Distribution of covariates x1 and x2
global mu0 = 70
global mu1 = 60
global sd0 = 10
global sd1 = 15
global px0 = 0.6


*************************************************************
*  1. Create sub-program that simulates data for a single subject  *
*************************************************************

capture program drop wholesim
program define wholesim, rclass

* Generate x1 - binary covariate
local u = uniform()
local x1 = (`u' > $px0)

* Generate x2 - normally distributed conditional on value of x1
local x2 = ($mu0 + invnorm(uniform())*$sd0)*(`x1' == 0) ///
         + ($mu1 + invnorm(uniform())*$sd1)*(`x1' == 1)

* Generate pscore - propensity score, a function of x1 and x2
local lpt = $a0 + $a1*`x1' + $a2*`x2'
local pscore = exp(`lpt')/(1 + exp(`lpt'))
```

```
* Generate z - treatment indicator, with P(z = 1) = pscore
local u = uniform()
local z = (‘u’ <= ‘pscore’)

* Generate y (outcome variable) - a function of z and x2
local y = $g0 + $g2*‘x2’ + $b0*‘z’ + invnorm(uniform())*$sde

* Store parameters of interest
return scalar y       = ‘y’
return scalar z       = ‘z’
return scalar pscore  = ‘pscore’
return scalar x1      = ‘x1’
return scalar x2      = ‘x2’

end


*************************************************************
*  2. Create sub-program that simulates a dataset of size 2,000   *
*************************************************************

capture program drop wholesim2
program define wholesim2, rclass

* Simulate a sample of size 2,000
simulate "wholesim" y=r(y) z=r(z) pscore=r(pscore) ///
x1=r(x1) x2=r(x2), reps(2000)

* Estimate propensity score using a logistic regression model
glm z x1 x2, fam(bin)
predict estpscore

* Define 2 strata by ‘true’ propensity score
xtile strata=pscore, nquantiles(2)

* Define 2 strata by estimated propensity score
xtile eststrata=estpscore, nquantiles(2)

* Estimate beta^s from the simulated dataset
forvalues s = 1 (1) 2 {
forvalues t = 0 (1) 1 {
summ y if z==‘t’ & strata==‘s’
local y‘t’‘s’ = r(mean)
summ y if z==‘t’ & eststrata==‘s’
local esty‘t’‘s’ = r(mean)
}
}
```

```
local beta    = 0
local estbeta = 0
forvalues s = 1 (1) 2 {
local beta    = `beta'    + (`y1`s''    - `y0`s'')/2
local estbeta = `estbeta' + (`esty1`s'' - `esty0`s'')/2
}

* Store parameters of interest
return scalar beta    = `beta'
return scalar estbeta = `estbeta'

end


*************************************************
*  3. Simulate 3,000 datasets of size 2,000   *
*************************************************

simulate "wholesim2"  beta=r(beta) estbeta=r(estbeta), ///
reps(3000)
```

This program creates a dataset containing two variables — **beta** and **estbeta**. The variable **beta** contains 3,000 stratified treatment effect estimates, β^s, that were obtained by stratifying on the true propensity score. The variable **estbeta** contains 3,000 stratified treatment effect estimates, $\hat{\beta}^s$, that were obtained by stratifying on an estimated propensity score, obtained by fitting a correctly specified maximum likelihood logistic regression model. The variance of the 3,000 values of **beta** is an empirical estimate of $\mathbb{V}_k[\,\hat{\beta}^s\,]$ and the variance of the 3,000 values of **estbeta** is an empirical estimate of $\mathbb{V}_e[\,\hat{\beta}^s\,]$.

D.2 Mathematica program used to obtain the population strata boundaries

```
************************************
* 0. Parameters of the problem     *
************************************

* Propensity score parameters
Clear[a0, a1, a2];
a0 = -3;
a1 = -0.5;
a2 = 0.05;

* Distribution of covariates x1 and x2
Clear[mu0, mu1, sd0, sd1, px0, px1];
mu0 = 70;    mu1 = 60;
sd0 = 10;    sd1 = 15;
px0 = 0.6;   px1 = 0.4;


****************************************
* 1. The p.d.f of the propensity score  *
****************************************

* (See Section C.2.1.)

Clear[v0, v1, w0, w1, f];

w0[p_] := (Log[p/(1 - p)] - a0)/a2;
w1[p_] := (Log[p/(1 - p)] - a0 - a1)/a2;
v0[p_] := Exp[-((w0[p] - mu0)^2)/(2*sd0^2)]/Sqrt[2*Pi*sd0^2];
v1[p_] := Exp[-((w1[p] - mu1)^2)/(2*sd1^2)]/Sqrt[2*Pi*sd1^2];

f[p_]  := (px0*v0[p] + px1*v1[p])/(a2*p*(1 - p));


****************************************
* 2. Obtaining the strata boundaries  *
****************************************

* (See Section C.2.2.)

* The cumulative distribution function of the propensity score, F(p)
Clear[F];
F[q_] := 0 /; q < 0
F[q_] := NIntegrate[f[p], {p, 0, q}] /; q <= 1 && q >= 0
F[q_] := NIntegrate[f[p], {p, 0, 1}] /; q > 1
```

```
Clear[psi];
psi[q_] := F[q] - 1/2;

* The population strata boundary solves psi[q] = 0.
* We need to suggest starting values for the solution.
FindRoot[psi[q] == 0, {q,  0.55, 0.551}];
```

D.3 Mathematica program used to obtain theoretical values of the variances

```
************************************
* 0. Parameters of the problem     *
************************************

* Population strata boundaries
Clear[q0, q1, q2]
q0 = 0;
q1 = 0.552393598869689 2;
q2 = 1;

* Propensity score parameters
Clear[a0, a1, a2];
a0 = -3;
a1 = -0.5;
a2 = 0.05;

* Individual-level treatment effect
Clear[b0];
b0 = 2;

* Outcome parameters
Clear[g0, g2, sde];
g0  = 35;
g2  = 0.3;
sde = 10;

* Distribution of covariates x1 and x2
Clear[mu0, mu1, sd0, sd1, px0, px1];
mu0 = 70;     sd0 = 10;
mu1 = 60;     sd1 = 15;
px0 = 0.6;    px1 = 0.4;

*******************************************
* 1. The p.d.f of the propensity score  *
*******************************************

* (See Section C.2.1.)

Clear[v0, v1, w0, w1, f];

w0[p_]:= (Log[p/(1 - p)] - a0)/a2;
w1[p_]:= (Log[p/(1 - p)] - a0 - a1)/a2;
v0[p_]:= Exp[-((w0[p] - mu0)^2)/(2*sd0^2)]/Sqrt[2*Pi*sd0^2];
v1[p_]:= Exp[-((w1[p] - mu1)^2)/(2*sd1^2)]/Sqrt[2*Pi*sd1^2];

f[p_] := (px0*v0[p] + px1*v1[p])/(a2*p*(1 - p));
```

```
*******************************************************
* 2. The probability of being treated and in each stratum  *
*******************************************************
```

* *(See Section C.2.3.)*

```
Clear[d1, d0];

d1[ql_,qu_]:= NIntegrate[p*f[p], {p,ql,qu}];
d0[ql_,qu_]:= 0.5 - d1[ql,qu];


*******************************************************
* 3. The population stratified treatment effect  *
*******************************************************
```

* *(See Section C.2.4.)*

```
Clear[iEy1, IEy1, Ey1, iEy0, IEy0, Ey0];

iEy1[p_]      := (px0*v0[p]*w0[p] + px1*v1[p]*w1[p])/(a2*(1 - p));
IEy1[ql_,qu_]:= NIntegrate[iEy1[p], {p,ql,qu}];
Ey1[ql_,qu_] := g0 + b0 + (g2/d1[ql,qu])*IEy1[ql,qu];

iEy0[p_]      := (px0*v0[p]*w0[p] + px1*v1[p]*w1[p])/(a2*p);
IEy0[ql_,qu_]:= NIntegrate[iEy0[p], {p,ql,qu}];
Ey0[ql_,qu_] := g0 + (g2/d0[ql,qu])*IEy0[ql,qu];

Clear[bs,betas];
bs[ql_,qu_]:= Ey1[ql,qu] - Ey0[ql,qu];

betas = (bs[q0,q1] + bs[q1,q2])/2;


*******************************************************
* 4. The conditional variance given treatment and strata  *
*******************************************************
```

* *(See Section C.2.5.)*

```
* The conditional variance given treatment and in each strata
Clear[iA1, IA1, A1, iB1, IB1, B1, Vy1];

iA1[p_]      := (px0*v0[p]*w0[p]^2 + px1*v1[p]*w1[p]^2)/(a2*(1 - p));
IA1[ql_,qu_]:= NIntegrate[iA1[p], {p, ql,qu}];
A1[ql_,qu_] := (g2^2/d1[ql,qu])*IA1[ql,qu];
```

```
iB1[p_]      := (px0*v0[p]*w0[p] + px1*v1[p]*w1[p])/(a2*(1 - p));
IB1[ql_,qu_]:= NIntegrate[iB1[p], {p,ql,qu}];
B1[ql_,qu_] := (g2/d1[ql,qu])*IB1[ql,qu];

Vy1[ql_,qu_]:= sde^2 + A1[ql,qu] - B1[ql,qu]^2;


* The conditional variance given no treatment and in each strata
Clear[iA0, IA0, A0, iB0, IB0, B0, Vy0];

iA0[p_]      := (px0*v0[p]*w0[p]^2 + px1*v1[p]*w1[p]^2)/(a2*p);
IA0[ql_,qu_]:= NIntegrate[iA0[p], {p,ql,qu}];
A0[ql_,qu_] := (g2^2/d0[ql,qu])*IA0[ql,qu];

iB0[p_]      := (px0*v0[p]*w0[p] + px1*v1[p]*w1[p])/(a2*p);
IB0[ql_,qu_]:= NIntegrate[iB0[p], {p,ql,qu}];
B0[ql_,qu_] := (g2/d0[ql,qu])*IB0[ql,qu];

Vy0[ql_,qu_]:= sde^2 + A0[ql,qu] - B0[ql,qu]^2;


*****************************************
* 5. The first variance component, n.V1  *
*****************************************
* (See Theorem 3.1.)

Clear[V1, matrixV1];

V1[ql_,qu_]:= 0.5^2(Vy1[ql,qu]/d1[ql,qu] + Vy0[ql,qu]/d0[ql,qu]);

matrixV1 = V1[q0,q1] + V1[q1,q2];


*********************************************************
* 6. The derivative of beta^* w.r.t. the strata boundaries  *
*********************************************************
* (See Section C.2.6.)

Clear[iF1, iG1, G1, dqEy1, iF0, iG0, G0, dqEy0];

iF1[p_]      := (px0*v0[p] + px1*v1[p])/(a2*(1 - p));
iG1[p_]      := (px0*v0[p]*w0[p] + px1*v1[p]*w1[p])/(a2*(1 - p));
G1[ql_,qu_]:= NIntegrate[iG1[p], {p,ql,qu}];

iF0[p_]      := (px0*v0[p] + px1*v1[p])/(a2*p);
iG0[p_]      := (px0*v0[p]*w0[p] + px1*v1[p]*w1[p])/(a2*p);
G0[ql_,qu_]:= NIntegrate[iG0[p], {p,ql,qu}];
```

```
dqEy1[qk_,ql_,qu_]:= (g2/d1[ql,qu])*iG1[qk] ///
                     - (g2*G1[ql,qu]/d1[ql,qu]^2)*iF1[qk];
dqEy0[qk_,ql_,qu_]:= (g2/d0[ql,qu])*iG0[qk] ///
                     - (g2*F0[ql,qu]/d0[ql,qu]^2)*iF0[qk];
```

* Then $\mathbf{dqEy[q_s,\ q_{(s-1)}, q_s]} = \frac{\partial}{\partial q_s}\{\mathbb{E}[Y \mid Z=1,\ S_s=1]\}$

* and $-\mathbf{dqEy[q_{(s-1)}, q_{(s-1)}, q_s]} = \frac{\partial}{\partial q_{s-1}}\{\mathbb{E}[Y \mid Z=1,\ S_s=1]\}$

```
Clear[dqbeta];
dqbeta[1]:=  0.5*(dqEy1[q1,q0,q1] - dqEy1[q1,q1,q2] ///
                - dqEy0[q1,q0,q1] + dqEy0[q1,q1,q2]);

matrixdqbeta = dqbeta[1];

*******************************************
* 7. The second variance component - n.V2 *
*******************************************
```

* *(See Theorem 3.1.)*

```
Clear[matrixncovq, matrixV2];
matrixncovq = (0.5*0.5)/(f[q1]^2);
matrixV2 = matrixdqbeta.matrixncovq.matrixdqbeta;


*******************************************
* 8. The covariance matrix of the estimated *
*    propensity score parameters            *
*******************************************
```

* *(See Section C.3.1.)*

```
Clear[iepx0, iepx1, iepx2, iepx1x2, iepx2x2]
iepx0[p_]  := (px0*v0[p] + px1*v1[p])/a2;
iepx1[p_]  :=  px1*v1[p]/a2;
iepx2[p_]  := (px0*v0[p]*w0[p] + px1*v1[p]*w1[p])/a2;
iepx1x2[p_]:=  px1*v1[p]*w1[p]/a2;
iepx2x2[p_]:= (px0*v0[p]*w0[p]^2 + px1*v1[p]*w1[p]^2)/a2;

Clear[epx0, epx1, epx2, epx1x2, epx2x2];
epx0   = NIntegrate[iepx0[p],   {p,0,1}];
epx1   = NIntegrate[iepx1[p],   {p,0,1}];
epx2   = NIntegrate[iepx2[p],   {p,0,1}];
epx1x2 = NIntegrate[iepx1x2[p], {p,0,1}];
epx2x2 = NIntegrate[iepx2x2[p], {p,0,1}];
```

```
Clear[matrixinvalpha, matrixncovalpha];
matrixinvalpha  = {{epx0, epx1,   epx2},
                   {epx1, epx1,   epx1x2},
                   {epx2, epx1x2, epx2x2}};

matrixncovalpha = Inverse[matrixinvalpha];


*****************************************
* 9. The covariance of outcomes, covariates  *
*    and the propensity score                *
*****************************************
```

(See Section C.3.2.)

```
* E[ X_k.(1 - p) | Z = 1, S_s = 1] and E[ X_k.p | Z = 0, S_s = 1]

Clear[iExp, Exp1, Exp0];
iExp[0,p_]:= (px0*v0[p] + px1*v1[p])/a2;
iExp[1,p_]:=  px1*v1[p]/a2;
iExp[2,p_]:= (px0*v0[p]*w0[p] + px1*v1[p]*w1[p])/a2;

Exp1[k_,ql_,qu_]:= NIntegrate[iExp[k,p], {p,ql,qu}]/d1[ql,qu];
Exp0[k_,ql_,qu_]:= NIntegrate[iExp[k,p], {p,ql,qu}]/d0[ql,qu];

* E[ Y.X_k.(1 - p) | Z = 1, S_s = 1] and E[ Y.X_k.p | Z = 0, S_s = 1]

Clear[h, iEyxp1, iEyxp0, Eyxp1, Eyxp0];
h[t_, r_, p_]:= g0 + g2*w1[p]*r + g2*w0[p]*(1 - r) + b0*t;

iEyxp1[0,p_]:= (px0*v0[p]*h[1,0,p] + px1*v1[p]*h[1,1,p])/a2;
iEyxp1[1,p_]:=  px1*v1[p]*h[1,1,p]/a2;
iEyxp1[2,p_]:= (px0*v0[p]*w0[p]*h[1,0,p] + px1*v1[p]*w1[p]*h[1,1,p])/a2;

iEyxp0[0,p_]:= (px0*v0[p]*h[0,0,p] + px1*v1[p]*h[0,1,p])/a2;
iEyxp0[1,p_]:=  px1*v1[p]*h[0,1,p]/a2;
iEyxp0[2,p_]:= (px0*v0[p]*w0[p]*h[0,0,p] + px1*v1[p]*w1[p]*h[0,1,p])/a2;

Eyxp1[k_,ql_,qu_]:= NIntegrate[iEyxp1[k,p], {p,ql,qu}]/d1[ql,qu];
Eyxp0[k_,ql_,qu_]:= NIntegrate[iEyxp0[k,p], {p,ql,qu}]/d0[ql,qu];


* Cov[Y, X_k.(1 - p) | Z = 1, S_s = 1] and Cov[Y, X_k.p | Z = 0, S_s = 1]

Clear[Cyxp1, Cyxp0];
Cyxp1[k_,ql_,qu_]:= Eyxp1[k,ql,qu]  - Exp1[k,ql,qu]*Ey1[ql,qu];
Cyxp0[k_,ql_,qu_]:= Eyxp0[k,ql,qu]  - Exp0[k,ql,qu]*Ey0[ql,qu];
```

```
Clear[matrixC, Cov1, Cov0, Cov];

Cov1[k_]:= Cyxp1[k,q0,q1] + Cyxp1[k,q1,q2];
Cov0[k_]:= Cyxp0[k,q0,q1] + Cyxp0[k,q1,q2];
Cov[k_] := (Cov1[k] + Cov0[k])/2;

matrixC = {Cov[0], Cov[1], Cov[2]};
```

```
*****************************************
* 10. The third variance component - n.V3 *
*****************************************

* (See Theorem 3.2.)

Clear[matrixV3];

matrixV3 = - matrixC.matrixncovalpha.matrixC;


*************************************************
* 11. Derivatives of functions v, w and t, w.r.t. alpha *
*************************************************

* Derivatives of w0, w1 w.r.t. alpha

Clear[daw0, daw1];
daw0[0,p_]:= - 1/a2;
daw0[1,p_]:=   0;
daw0[2,p_]:= - w0[p]/a2;

daw1[0,p_]:= - 1/a2;
daw1[1,p_]:= - 1/a2;
daw1[2,p_]:= - w1[p]/a2;

* Derivatives of v0 and v1 w.r.t. alpha

Clear[dav0, dav1];
dav0[0,p_]:= v0[p]*(w0[p] - mu0)/(a2*sd0^2);
dav0[1,p_]:= 0;
dav0[2,p_]:= v0[p]*w0[p]*(w0[p] - mu0)/(a2*sd0^2)

dav1[0,p_]:= v1[p]*(w1[p] - mu1)/(a2*sd1^2);
dav1[1,p_]:= v1[p]*(w1[p] - mu1)/(a2*sd1^2);
dav1[2,p_]:= v1[p]*w1[p]*(w1[p] - mu1)/(a2*sd1^2);
```

```
* Derivatives of t0, t1 w.r.t. alpha

Clear[dat0, dat1];
dat0[0,p_]:= - p*(1 - p);
dat0[1,p_]:=   0;
dat0[2,p_]:= - p*(1 - p)*w0[p];

dat1[0,p_]:= - p*(1 - p);
dat1[1,p_]:= - p*(1 - p);
dat1[2,p_]:= - p*(1 - p)*w1[p];


*************************************************
* 12.  The derivative of the c.d.f, F(p), of the *
*      propensity score, w.r.t. alpha.           *
*************************************************
```

* *(See Section C.3.3.)*

```
Clear[daf, daF];
daf[k_,p_]:= (px0*daw0[k,p] + px1*daw1[k,p])/(a2*p*(1 - p));
daf[2,p_] := (px0*daw0[2,p] + px1*daw1[2,p])/(a2*p*(1 - p)) ///
             - (px0*w0[p] + px1*w1[p])/(a2^2*p*(1 - p));
daF[k_]:=  NIntegrate[daf[k,p], {p,0,q1}];


*************************************************
* 13. The integrals I_f1k, I_f0k, I_Y1k and I_Y0k *
*************************************************
```

(See Section C.3.4.)

```
(* The integrals I_f1k and I_f0k *)
Clear[Ipdaf1, Ipdaf0];
If1[k_,ql_,qu_]:= NIntegrate[      p*daf[k,p], {p,ql,qu}];
If0[k_,ql_,qu_]:= NIntegrate[(1 - p)*daf[k,p], {p,ql,qu}]


(* The conditional expectations  E[Y_1 | p]  and  E[Y_0 | p] *)
Clear[Eypa, Eypb, Eyp];
Eypa[t_,p_]:= (px0*v0[p]*h[t,0,p])/(px0*v0[p] + px1*v1[p]);
Eypb[t_,p_]:= (px1*v1[p]*h[t,1,p])/(px0*v0[p] + px1*v1[p]);
Eyp[t_,p_] := Eypa[t,p] + Eypb[t,p];
```

```
(* The derivative of   E[ Y_1 | p(x; alpha) = r ]    w.r.t. alpha *)
Clear[daEypa, daEypb, daEypc, daEypd, daEype, daEypf,  daEyp];
daEypa[t_,k_,p_]:=  (px0*dav0[k,p]*h[t,0,p])/(px0*v0[p] + px1*v1[p]);
daEypb[t_,k_,p_]:=  (px0*v0[p]*g2*daw0[k,p])/(px0*v0[p] + px1*v1[p]);
daEypc[t_,k_,p_]:= -(px0*v0[p]*h[t,0,p])/(px0*v0[p] + px1*v1[p])^2;

daEypd[t_,k_,p_]:=  (px1*dav1[k,p]*h[t,1,p])/(px0*v0[p] + px1*v1[p]);
daEype[t_,k_,p_]:=  (px1*v1[p]*g2*daw1[k,p])/(px0*v0[p] + px1*v1[p]);
daEypf[t_,k_,p_]:= -(px1*v1[p]*h[t,1,p])/(px0*v0[p] + px1*v1[p])^2;

daEyp[t_,k_,p_]:= daEypa[t,k,p] + daEypb[t,k,p] + daEypc[t,k,p] ///
                  daEypd[t,k,p] + daEype[t,k,p] + daEypf[t,k,p];


(* The integral of  p.f_p(r; alpha).d/d alpha {E[Y| p(x;alpha) = r]} *)
Clear[IfY1, IfY0];
IfY1[k_,ql_,qu_]:= NIntegrate[      p*f[p]*daEyp[1,k,p], {p,ql,qu}];
IfY0[k_,ql_,qu_]:= NIntegrate[(1 - p)*f[p]*daEyp[0,k,p], {p,ql,qu}];


(* The integral of  p.E[Y | p(x;alpha) = r].d/d alpha{f_p(r;alpha)} *)
Clear[IY1f, IY0f];
IY1f[k_,ql_,qu_]:= NIntegrate[      p*Eyp[1,p]*daf[k,p], {p,ql,qu}];
IY0f[k_,ql_,qu_]:= NIntegrate[(1 - p)*Eyp[0,p]*daf[k,p], {p,ql,qu}];


(* The integral of p.d/d alpha {E[Y | p(x;alpha) = r].f_p(r; alpha)} *)
Clear[IpdaEypf1, IpdaEypf0];
IY1[k_, ql_, qu_] := IY1f[k, ql, qu] + IfY1[k, ql, qu];
IY0[k_, ql_, qu_] := IY0f[k, ql, qu] + IfY0[k, ql, qu];


**********************************************
* 14. The fourth variance component - n.V4 *
**********************************************
```

* *(See Theorem 3.2.)*

```
(* The vector e *)
Clear[ea1s, ea0s, ea1, ea0, ea, eq, e, matrixD];
ea1s[k_,ql_,qu_]:= (IY1[k,ql,qu] - Ey1[ql,qu]*If1[k,ql,qu])/(2*d1[ql, qu]);
ea0s[k_,ql_,qu_]:= (IY0[k,ql,qu] - Ey0[ql,qu]*If0[k,ql,qu])/(2*d0[ql, qu]);

ea1[k_]:= ea1s[k,q0,q1] + ea1s[k,q1,q2] + ea1s[k,q2,q3] ///
        + ea1s[k,q3,q4] + ea1s[k,q4,q1];
ea0[k_]:= ea0s[k,q0,q1] + ea0s[k,q1,q2] + ea0s[k,q2,q3] ///
        + ea0s[k,q3,q4] + ea0s[k,q4,q1];
ea[k_]:=  ea1[k] - ea0[k];
```

```
eq[k_] := dqbeta[1]*daF[k]/f[q1];
e[k_] := ea[k] - eq[k];
matrixE = {e[0], e[1], e[2]};


Clear[matrixV4];

matrixV4 = matrixE.matrixncovalpha.matrixE;
```

```
*************************************************************
* 15. The theoretical variances, with the propensity score *
*     known and estimated                                   *
*************************************************************

Clear[Vkbeta, Vebeta];
Vkbeta = matrixV1 + matrixV2;
Vebeta = matrixV1 + matrixV2 + matrixV3 + matrixV4;
```

References

[1] A. Abadie and G. Imbens. Simple and bias-corrected matching estimators for average treatment effects. *NBER Working Paper*, available at http://citeseer.ist.psu.edu/abadie02simple.html, (2002).

[2] R. Agodini and M. Dynarski. Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics*, 86:180–194, (2004).

[3] K. Anstrom and A. Tsiatis. Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data. *Biometrics*, 57:1207–1218, (2001).

[4] P. Austin and M. Mamdani. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine*, 25:2084–2106, (2006).

[5] B. Barnow. The impact of CETA programs on earnings: a review of the literature. *Journal of Human Resources*, 22:157–193, (1987).

[6] R. Beals. *Analysis: an introduction.* Cambridge: Cambridge University Press, 2004.

[7] G. Beaumont. *Elementary mathematical statistics.* London: McGraw-Hill, (1972).

[8] N. Bellamy, W. Buchanan, C. Goldsmith, J. Campbell, and L. Stitt. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *Journal of Rheumatology*, 15:1833–1840, (1988).

[9] K. Benson and A. Hartz. A comparison of observational studies and randomized. controlled trials. *New England Journal of Medicine*, 342:1878–1886, (2000).

[10] N. Black. Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal*, 312:1215–1218, (1996).

[11] L. Braitman and P. Rosenbaum. Rare outcomes, common treatments: analytic strategies using propensity scores [Editorial]. *Annals of Internal Medicine*, 137:693–695, (2002).

[12] M. Bulmer. *Principles of statistics.* New York: Dover Publications, (1979).

[13] J. Clinch, P. Tugwell, G. Wells, and B. Shea. Individualized functional priority approach to the assessment of health related quality of life in rheumatology. *Journal of Rheumatology*, 28:445–451, (2001).

[14] W. Cochran. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24:295–313, (1968).

[15] J. Concato, N. Shah, and R. Horwitz. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, 342:1887–1892, (2000).

[16] D. Cox. Note on grouping. *Journal of the American Statistical Association*, 52:543–547, (1957).

[17] D. Cox and D. Hinkley. *Theoretical statistics.* London: Chapman & Hall, (1974).

[18] R. D'Agostino. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17:2265–2281, (1998).

[19] R. D'Agostino and D. Rubin. Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 95:749–759, (2000).

[20] A. Davison and D. Hinkley. *Bootstrap methods and their application.* Cambridge: Cambridge University Press, 1997.

[21] A. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society B*, 41:1–31, (1979).

[22] R. Dehejia. Practical propensity score matching: a reply to Smith and Todd. *Journal of Econometrics*, 125:353–364, (2005).

[23] R. Dehejia and S. Wahba. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94:1053–1062, (1999).

[24] C. Drake. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 49:1231–1236, (1993).

[25] S. Emerson and J. Emerson. Direct standardization of incidence rates in the presence of incomplete data. *Statistics in Medicine*, 12:3–12, (1993).

[26] D. Friedlander and P. Robins. Evaluating program evaluations: new evidence on commonly used nonexperimental methods. *American Economic Review*, 85:923–937, (1995).

[27] M. Frölich. Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics*, 86:77–90, (2004).

[28] M. Gecht, K. Connell, J. Sinacore, and T. Prohaska. A survey of exercise beliefs and exercise habits among people with arthritis. *Arthritis Care and Research*, 9:82–88, (1996).

[29] M. Giurcanu and A. Trinidade. Establishing consistency of M-estimators under concavity with an application to some financial risk measures. *Technical Report 2005-024, Department of Statistics, University of Florida*, available at http://www.stat.ufl.edu/ giurcanu/papers.html, (2005).

[30] P. Green and B. Silverman. *Nonparametric regression and generalized linear models.* London: Chapman & Hall, (1994).

[31] S. Greenland and B. Brumback. An overview of relations among causal modelling methods. *International Journal of Epidemiology*, 31:1030–1037, (2002).

[32] X. Gu and P. Rosenbaum. Comparison of multivariate matching methods: structures, distances and algorithms. *Journal of Computational and Graphical Statistics*, 2:405–420, (1993).

[33] A. Hammond and N. Lincoln. The effect of a joint protection education programme for people with rheumatoid arthritis. *Clicical Rehabilitation*, 13:392–400, (1999).

[34] W. Härdle. *Applied nonparametric regression.* Cambridge: Cambridge University Press, (1990).

[35] D. Hawley. Psycho-educational interventions in the treatment of arthritis. *Balliere's Clinical Rheumatology*, 9:803–823, (1995).

[36] J. Heckman, H. Ichimura, J. Smith, and P. Todd. Characterizing selection bias using experimental data. *Econometrica*, 66:1017–1098, (1998).

[37] M. Hernán. A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health*, 58:265–271, (2004).

[38] J. Hill and J. Reiter. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, 25:2230–2256, (2006).

[39] K. Hirano and G. Imbens. The propensity score with continuous treatments. *Draft of a chapter for 'Missing Data and Bayesian Methods in Practice: Contributions by Donaly Rubin's Statistical Family'*, available at http://elsa.berkeley.edu/∼imbens/hir$_0$7feb04.pdf, (forthcoming from Wiley).

[40] K. Hirano, G. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71:1161–1189, (2003).

[41] P. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960, (1986).

[42] G. Hong and S. Raudenbush. Evaluating kindergarten retention policy: a case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101:901–910, (2006).

[43] P. Huber. *Robust statistics.* New York: John Wiley & Sons, (1981).

[44] K. Hullsiek and T. Louis. Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics*, 3:179–193, (2002).

[45] M. Hurley, N. Walsh, H. Mitchell, T. Pimm, A. Patel, E. Williamson, R. Jones. P. Dieppe, and B. Reeves. Clinical effectiveness of ESCAPE - knee pain a rehabilitation programme for chronic knee pain: A cluster randomised trial. *Arthritis Care and Research*, (in press).

[46] G. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87:706–710, (2000).

[47] G. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86:4–29, (2004).

[48] J. Ioannidis, A. Haidich, and J. Lau. Any casualties in the clash of randomised and observational evidence? *British Medical Journal*, 322:879–880, (2001).

[49] M. Joffe and P. Rosenbaum. Invited commentary: propensity scores. *American Journal of Epidemiology*, 150:327–333, (1999).

[50] D. Judkins, D. Morganstein, P. Zador, A. Piesse, B. Barrett, and P. Mukhopadhyay. Variable selection and raking in propensity scoring. *Statistics in Medicine*, available at http:
www3.interscience.wiley.com/cgi-bin/jissue/96515927, (in press).

[51] P. June, L. Nartey, S. Reichenbach, R. Sterchi, P. Dieppe, and M. Egger. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *American Journal of Epidemiology*, 364:2021–2029, (2004).

[52] B. Kirkwood and J. Sterne. *Essential medical statistics, 2^{nd} edition.* Malden: Blackwell Science, (2003).

[53] M. Knapen, P. Zusterzeel, W. Peters, E. Steegers, J. Lefebvre, F. Keefe, G. Affleck, L. Raezer, K. Starr, D. Caldwell, and H. Tennen. The relationship of arthritis self-efficacy to daily pain, daily mood, and daily pain coping in rheumatoid arthritis patients. *Pain*, 80:425–435, (1999).

[54] R. Kunz and A. Oxman. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *British Medical Journal*, 317:1185–1190, (1998).

[55] T. Kurth, A. Walker, R. Glynn, K. Chan, J. Gaziano, K. Berger, and J. Robins. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*, 163:262–270, (2006).

[56] R. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76:604–620, (1986).

[57] R. Little and D. Rubin. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, 21:121–145, (2000).

[58] R. Little and D. Rubin. *Statistical analysis with missing data, 2nd edition.* New York: John Wiley & Sons, (2002).

[59] J. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23:2937–2960, (2004).

[60] C. MacLean, K. Knight, H. Paulus, R. Brook, and P. Shekelle. Costs attributable to osteoarthritis. *Journal of Rheumatology*, 25:2213–2218, (1998).

[61] S. Maliski, L. Kwan, T. Krupski, A. Fink, J. Orecklin, and M. Litwin. Confidence in the ability to communicate with physicians among low-income patients with prostate cancer. *Urology*, 64:329–334, (2004).

[62] N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22:719–748, (1959).

[63] J. Marron and D. Nolan. Canonical kernels for density estimation. *Statistics and Probability Letters*, 7:195–199, (1988).

[64] E. Martens, W. Pestman, A. de Boer, S. Belitser, and O. Klungel. Instrumental variables. Application and limitations. *Epidemiology*, 17:260–267, (2006).

[65] L. McCandless, P. Gustafson, and P. Austin. Bayesian propensity score analysis for observational data. *American Journal of Epidemiology*, 163: Suppl. S222, (2006).

[66] M. McKee, A. Britton, N. Black, K. McPherson, C. Sanderson, and C. Bain. Interpreting the evidence: choosing between randomised and non-randomised studies. *British Medical Journal*, 319:312–315, (1999).

[67] C. Michalopoulos, H. Bloom, and C. Hill. Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? *Review of Economics and Statistics*, 86:156–179, (2004).

[68] O. Miettinen. Matching and design efficiency in retrospective studies. *American Journal of Epidemiology*, 91:111–118, (1970).

[69] O. Miettinen and E. Cook. Confounding: essence and detection. *American Journal of Epidemiology*, 114:593–603, (1981).

[70] K. Ming and P. Rosenbaum. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*, 56:118–124, (2000).

[71] E. Nadaraya. On estimating regression. *Theory Prob. Appl.*, 10:186–190, (1964).

[72] E. Parzen. On estimation of a probability density and mode. *Annals of Mathematical Statistics*, 35:1065–1076, (1962).

[73] J. Pearl. *Causality*. Cambridge: Cambridge University Press, (2000).

[74] G. Peat, R. McCarney, and P. Croft. Knee pain and osteoarthritis in older adults: a review of community burden and current use of primary health care. *Annals of the Rheumatic Diseases*, 60:91–97, (2001).

[75] S. Pocock and D. Elbourne. Randomized trials or observational tribulations? *New England Journal of Medicine*, 342:1907–1909, (2000).

[76] J. Preisser, M. Young, D. Zaccaro, and M. Wolfson. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Statistics in Medicine*, 22:1235–1254, (2003).

[77] J. Robins and S. Greenland. The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology*, 123:392–402, (1986).

[78] J. Robins, M. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560, (2000).

[79] J. Robins, S. Mark, and W. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48:479–495, (1992).

[80] J. Robins, A. Rotnitzky, and L. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, (1994).

[81] P. Rosenbaum. Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79:565–571, (1984).

[82] P. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82:387–394, (1987).

[83] P. Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84:1024–1032, (1989).

[84] P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, (1983).

[85] P. Rosenbaum and D. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516–524, (1984).

[86] P. Rosenbaum and D. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, 39:33–38, (1985).

[87] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:642–669, (1956).

[88] D. Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29:185–203, (1973).

[89] D. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74:318–328, (1979).

[90] D. Rubin. Bias reduction using Mahalanobis-metric matching. *Biometrics*, 36:293–298, (1980).

[91] D. Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127:757–763, (1997).

[92] D. Rubin. On principles for modeling propensity scores in medical research [Editorial]. *Pharmacoepidemiology and Drug Safety*, 13:855–857, (2004).

[93] D. Rubin and N. Thomas. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95:573–585, (2000).

[94] M. SchrumpfHeiberg, E. Rodevand, K. Mikkelsen, C. Kaufmann, A. Didriksen, P. Mowinckel, and T. Kvien. Adalimumab plus methotrexate is more effective than adalimumab alone in patients with established rheumatoid arthritis: Results from a 6-month longitudinal, observational, multicenter study. *Annals of the Rheumatic Diseases*, 10:1379–1383, (2006).

[95] B. Shah, A. Laupacis, J. Hux, and P. Austin. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology*, 58:550–559. (2005).

[96] B. Silverman. *Density estimation for statistics and data analysis.* London: Chapman & Hall, (1986).

[97] H. Smith. Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology*, 27:325–353, (1997).

[98] J. Smith and P. Todd. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125:305–353, (2005).

[99] StataCorp. *Stata Statistical Software. Release 9.* College Station, TX, (2005).

[100] L. Stefanski and D. Boos. The calculus of M-estimation. *American Statistician*, 56:29–38, (2002).

[101] T. Stukel, E. Fisher, D. Wennberg, D. Alter, D. Gottlieb, and M. Vermeulen. Analysis of observational studies in the presence of treatment selection bias. *American Journal of the American Medical Association*, 297:278–285, (2007).

[102] T. Stürmer, M. Joshi, R. Glynn, J. Avorn, K. Rothman, and S. Schneeweiss. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, 59:437–447, (2006).

[103] T. Stürmer, S. Schneeweiss, M. Brookhart, K. Rothman, J. Avorn, and R. Glynn. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *American Journal of Epidemiology*, 161:891–898, (2005).

[104] T. Stürmer, S. Schneeweiss, and R. Glynn. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology*, 162:279–289, (2005).

[105] E. Süli and D. Mayers. *An introduction to numerical analysis.* Cambridge: Cambridge University Press, 2003.

[106] R. Tannen, M. Weiner, and S. Marcus. Simulation of the Syst-Eur randomized control trial using a primary care electronic medical record was feasible. *Journal of Clinical Epidemiology*, 59:254–264, (2006).

[107] W. Tu and X. Zhou. A bootstrap confidence interval procedure for the treatment effect using propensity score subclassification. *UW Biostatistics Working Paper Series*, available at http://www.bepress.com/uwbiostat/paper200, (2003).

[108] A. van der Vaart. *Asymptotic statistics.* Cambridge: Cambridge University Press, (1998).

[109] G. Watson. Smooth regression analysis. *Sankhya, Series A*, 26:101–116, (1964).

[110] C. Weinberg. Toward a clearer definition of confounding. *American Journal of Epidemiology*, 137:1–8, (1993).

[111] S. Weitzen, K. Lapane, A. Toledano, A. Hume, and V. Mor. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety*, 13:841–853, (2004).

[112] Wolfram Research, Inc. *Mathematica. Version 5.2.* Champaign, Illinois, (2005).

[113] E. Zanutto. A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science*, 4:67–91, (2006).

[114] Z. Zhao. Using matching to estimate treatment effects: data requirements, matching metrics and Monte Carlo evidence. *Review of Economics and Statistics*, 86:91–107, (2004).

[115] Z. Zhao. Sensitivity of propensity score methods to the specifications. *IZA Discussion paper No. 1873*, available at http://ssrn.com/abstract=869005, (2005).

[116] A. Zigmond and R. Snaith. The hospital anxiety and depression scale. *Acta Psychiatrica*, 67:361–367, (1983).